Scale length does matter: Recommendations for Measurement Invariance Testing with Categorical Factor Analysis and Item Response Theory Approaches

E. Damiano D'Urso, Kim De Roover, Jeroen K. Vermunt, Jesper Tijmstra

Tilbug University, The Netherlands

Abstract

In social sciences, the study of group differences concerning latent constructs is ubiquitous. These constructs are generally measured by means of scales composed of ordinal items. In order to compare these constructs across groups, one crucial requirement is that they are measured equivalently or, in technical jargon, that measurement invariance holds across the groups. This study compared the performance of multiple group categorical confirmatory factor analysis (MG-CCFA) and multiple group item response theory (MG-IRT) in testing measurement invariance with ordinal data. A simulation study was conducted to compare the true positive rate (TPR) and false positive rate (FPR) both at the scale and at the item level for these two approaches under an invariance and a non-invariance scenario. The results of the simulation studies showed that the performance, in terms of the TPR, of MG-CCFA- and MG-IRT-based approaches mostly depends on the scale length. In fact, for long scales, the likelihood ratio test (LRT) approach, for MG-IRT, outperformed the other approaches, while, for short scales, MG-CCFA seemed to be generally preferable. In addition, the performance of MG-CCFA's fit measures, such as RMSEA and CFI, seemed to depend largely on the length of the scale, especially when MI was tested at the item level. General caution is recommended when using these measures, especially when MI is tested for each item individually. A decision flowchart, based on the results of the simulation studies, is provided to help summarizing the results and providing indications on which approach performed best and in which setting.

Scale length does matter: Recommendations for Measurement Invariance Testing with Categorical Factor Analysis and Item Response Theory Approaches

1 Introduction

One of the main missions of psychological and social sciences is to study individuals as well as group differences with regard to latent constructs (e.g., extraversion). Such constructs are commonly measured by means of psychological scales in which subjects rate their level of agreement on various Likert-scale type of items by selecting one out of the possible response options. Most items' response options range from 3 to 5 with a clear ordering (e.g., a score of 3 is higher than a score of 2 which is then higher than 1). Such items with few naturally ordered categories are called ordinal items.

In order to compare psychological constructs across groups one crucial requirement is that the construct is measured equivalently (Borsboom, 2006; Meredith & Teresi, 2006). Equivalence in the measurement of a psychological construct across groups is generally defined as measurement invariance (MI).

Latent variable modeling is one of the most popular frameworks in the context of psychological measurement. Within this framework various approaches have been developed to model ordinal data as well as to test for MI. Among them, two of the most used ones are multiple group categorical confirmatory factor analysis (MG-CCFA) and multiple group item response theory (MG-IRT)(E. S. Kim & Yoon, 2011; Millsap, 2012). These two approaches have been commonly seen as separate but they show quite some overlap in the context of ordinal data. For instance, parameters in MG-CCFA and MG-IRT models are known to be directly related (Takane & De Leeuw, 1987). Moreover, Chang, Hsu, and Tsai (2017) proposed a set of minimal identification constraints to make MG-CCFA and MG-IRT models fully equivalent.

The equivalence of these models does not necessarily match the way MI is conceptualized and tested within each of the two approaches. For example, one main difference between MG-CCFA and MG-IRT pertains to which hypotheses are tested. On the one hand, in MG-CCFA equivalence of measurement is mainly investigated at the scale level. That is, the tested hypothesis is that the complete set of items functions equivalently across

groups. On the other hand, in MG-IRT more attention is dedicated toward the study of each individual item. For this reason, MI is tested for each item in the scale separately. Another crucial difference is in the way these hypotheses are tested. In fact, different testing strategies are used to determine whether MI holds, either at the scale or at the item level.

The impact that these differences have in terms of the performance to detect MI has not yet been thoroughly assessed. Previous studies focused either on comparing MG-CCFA and MG-IRT using different models (Meade & Lautenschlager, 2004; E. S. Kim & Yoon, 2011) or, in the case of equivalent models, by solely using an item-level testing perspective (Chang et al., 2017). Providing clear guidelines on which approach to choose and in which setting is particularly helpful for applied researchers. In fact, it might facilitate decisions regarding the level at which (non)invariance will be tested (e.g., scale or item level) as well as what the most powerful tools to test it are. However, in the current literature, clear guidelines have not been yet provided. In this paper, by means of two simulation studies, we bring three main contributions: (i) assess to what extent performing a scale- or an item-level test affects the power to detect MI, (ii) determine what MG-CCFA- or MG-IRT-based testing strategies/measures are more powerful to test MI, and (iii) based on the results of the simulation studies, provide guidelines on what approach to choose and in which conditions.

To this end, in Section 2 will discuss both MG-CCFA- and MG-IRT-based models and illustrate how they are equivalent under a set of minimal identification constraints. Additionally, in the same section, for each of the two approaches, we will discuss the differences in the set of hypotheses and the testing strategies in the context of MI. Afterwards, in Section 3 we will assess the performance of MG-CCFA- and MG-IRT-based testing strategies in testing MI by means of two simulation studies. Finally, in Section 4 we will conclude by giving remarks and recommendations along with a summary of the main results obtained in the simulation studies.

2 MG-CCFA, MG-IRT models and their MI test

2.1 The models

Imagine to have data composed of J items for a group of N subjects. Also, assume that a grouping variable exists such that subjects can be divided in G. Let X_j be the response on item j and further assume that X_j is a polytomously scored response which might take on C possible values, with $c = \{0,1,2,...,C\text{-}1\}$. Without loss of generality, it can be assumed that a unidimensional construct η underlies the observed responses.

2.1.1 MG-CCFA. In MG-CCFA it is assumed that C possible observed values are obtained from a discretization of a continuous unobserved response variable X_j^* via some threshold parameters. The threshold $\tau_{j,c}^{(g)}$ indicates the dividing point for the categories (e.g., division between a score of 3 and 4). Additionally, they are created such that the first and the last threshold are defined as $\tau_{j,0}^{(g)} = \text{-}\infty$ and $\tau_{j,C}^{(g)} = +\infty$, respectively. Rewriting formally what we just described, we have:

$$X_j = c, \quad if \quad \tau_{j,c}^{(g)} < X_j^* < \tau_{j,c+1}^{(g)} \quad c = 0, 1, 2, ..., C - 1. \tag{1}$$

If it is also assumed that the construct under study is unidimensional, according to a factor analytical model we have:

$$X_j = \lambda_j^{(g)} \eta + \epsilon_j, \quad j = 1, 2, ..., J. \tag{2}$$

Equation (2) shows that the unobserved continuous response variable X_j^* is determined by a latent variable score η via the factor loading $\lambda_j^{(g)}$ and a residual component ϵ_j. The latter represents an error term that is item specific. It is important to note that the thresholds $\tau_{j,c}^{(g)}$ and loadings $\lambda_j^{(g)}$ are group specific. Additionally, within group g both the latent variable η and the item-specific residual component ϵ_j are mutually independent and both normally distributed, with:

$$\eta \sim N(\kappa, \varphi), \quad \text{and } \epsilon_j \sim N(0, \sigma_j^2). \tag{3}$$

where κ is the factor mean, φ the factor variance and σ_j^2 is the unique variance.

2.1.2 MG-noGRM. MG-IRT models the probability of selecting a specific item category, given a score on the latent construct and given a specific group membership. These conditional probabilities, in the case of ordinal items, are modeled indirectly through building blocks thar are constructed by means of specific functions. Different functions exist for ordinal items which, in turn, are used by different MG-IRT models. Because of its similarities with MG-CCFA (Chang et al., 2017), in the following, we only consider the multiple group normal ogive graded response model (MG-noGRM; Samejima, 1969). The MG-noGRM uses cumulative probabilities as its building blocks, and the underlying idea is to treat the multiple categories in a dichotomous fashion (Samejima, 1969). First, for every score the probability of obtaining that score or higher is calculated (e.g., selecting 2 or above), given the latent construct η. Based on this set of probabilities, the probability of selecting a specific category (e.g., 2) is calculated, given a certain score on η. In the MG-noGRM, like in MG-CCFA, it is assumed that the observed values arise from an underlying continuous latent response variable X_j^*.

Rewriting formally what we just described, the probability of scoring a certain category c is then:

$$\begin{aligned} P(X_j^* = c|\eta, g) &= \Phi(\alpha_j^{(g)}(\eta - \delta_{j,c+1}^{(g)})) - \Phi(\alpha_j^{(g)}(\eta - \delta_{j,c}^{(g)})) \\ &= \Phi(\alpha_j^{(g)}\eta - \alpha_j^{(g)}\delta_{j,c+1}^{(g)}) - \Phi(\alpha_j^{(g)}\eta - \alpha_j^{(g)}\delta_{j,c}^{(g)}) \\ &= \int_{\alpha_j^{(g)}\eta - \alpha_j^{(g)}\delta_{j,c}^{(g)}}^{\alpha_j^{(g)}\eta - \alpha_j^{(g)}\delta_{j,c+1}^{(g)}} \phi(u_j)du_j \end{aligned} \tag{4}$$

where, for group g $\alpha_j^{(g)}$ is the discrimination parameter for item j, and $\delta_{j,c}^{(g)}$ is the threshold parameter. The latter represents the point at which the probability of answering at or above category c is .5 for group g. Since ordered categories are modeled, the probability of getting at least the lowest score is 1, and the first threshold $\delta_{j,0}^{(g)}$ is not estimated and set to -∞. That is, C-1 threshold parameters per group need to be estimated. It is relevant to highlight that, like in MG-CCFA, also in the case of the MG-noGRM the model parameters $\alpha_j^{(g)}$ and $\delta_{j,c}^{(g)}$ are group specific. Also, ϕ is the cumulative density function for the standard normal variable u_j and $\Phi(u)$ is the cumulative distribution function.

2.1.2.1 Similarities with MG-CCFA. The similarities between MG-CCFA and the MG-noGRM can be revealed by taking a closer look at how the parameters in the two models are related (Takane & De Leeuw, 1987; Kamata & Bauer, 2008; Chang et al., 2017):

$$\alpha_j^{(g)} = \frac{\lambda_j^{(g)}}{\sigma_j}, \qquad u_j = \frac{\epsilon_j}{\sigma_j}, \qquad \delta_{j,c}^{(g)} = \frac{-\tau_{j,c}^{(g)}}{\alpha_j^{(g)}}, \tag{5}$$

and how it is possible to write the probability of X_j^* given η in MG-CCFA terms:

$$\begin{aligned} P(X_j^* = c|\eta, g) &= \int_{\lambda_j^{(g)}\eta - \tau_{j,c}^{(g)}}^{\lambda_j^{(g)}\eta - \tau_{j,c+1}^{(g)}} \phi(\epsilon_j) d\epsilon_j \\ &= \int_{\lambda_j^{(g)}\eta/\sigma_j - \tau_{j,c}^{(g)}/\sigma_j}^{\lambda_j^{(g)}\eta/\sigma_j - \tau_{j,c+1}^{(g)}/\sigma_j} \phi(u_j) du_j. \end{aligned} \tag{6}$$

The difference between (4) and (6) is that in MG-CCFA the loadings $\lambda_j^{(g)}$ and the thresholds $\tau_{j,c}^{(g)}$ can be inferred only in a relative sense. In fact, they can only be calculated through the ratio with the residual variance σ_j (Takane & De Leeuw, 1987; Kamata & Bauer, 2008; Chang et al., 2017). This is due to the absence of a scale for the latent response variable X_j^*. For ease of reading, in the following, only the term loading will be used to refer to both the discrimination parameters and the loadings.

2.1.3 Identification constraints and models equivalence. In order to identify measurement models such as the ones considered here, constraints are usually imposed either via specification of an arbitrary value for some parameters or by setting equalities across them. This way the number of parameters to be estimated is reduced, and it is possible to find a unique solution in the estimation process (Chang et al., 2017; San Martín & Rolin, 2013; Millsap & Yun-Tein, 2004).

In testing MI with multiple groups, both for MG-CCFA and the MG-noGRM, it is necessary to ensure that a scale is set for (i) the latent response variable X_j^*, (ii) the latent construct η, and that (iii) the scale of the latent construct is aligned across groups such that the parameters can be directly compared (Kamata & Bauer, 2008, Chang et al., 2017). Interestingly, these constraints are commonly imposed in a different way in MG-CCFA and in the MG-noGRM.

The observed response for each item is assumed to arise, in both models, from an unobserved continuous response variable X_j^*. These underlying continuous response variables

do not have a scale. For this reason, a scale has to be set by constraining their variances and means. In both models the means of the latent response variables are constrained to be 0. However, different ways to constrain the variances are generally used. It is common to either set their total variances $V(X_j^*)$ to 1 or its unique variances σ_j^2 to 1. The former is much more common in MG-CCFA while the latter is closer to what is usually done with the MG-noGRM (Kamata & Bauer, 2008). The other unobserved element for which a scale has to be set is the latent construct η. Again, this is commonly approached in a different way in the two frameworks. On the one hand, in MG-CCFA a fixed value is commonly chosen for a threshold and a loading. On the other hand, in the MG-noGRM the scale of the latent variable is commonly defined by setting its mean and variance to 0 and 1, respectively. In both cases these constraints are applied only for one of the two groups, which is usually called the reference group.

Finally, it is necessary to align the scale of both groups to make them comparable. This is commonly achieved by imposing equality constraints on some of the parameters in the model. Again, in MG-CCFA and in the MG-noGRM the common way to address this issue is different. On the one hand, in MG-CCFA for each latent construct, the factor loading and the threshold of a single item are constrained to be equal across groups. Generally, the loading and the threshold of the first item of the scale are selected. On the other hand, in MG-IRT multiple items, assumed to function equivalently in both groups, are set equal by constraining their parameters. These items form what is then called the anchor. Note that, in the MG-noGRM, and more generally in MG-IRT models, a bigger attention is devoted to choosing the items that are constrained to be equal across groups while in MG-CCFA this is not necessarily the case.

Chang et al. (2017) have recently proposed a set of minimal constraints to make MG-CCFA and the MG-noGRM fully comparable which will also be presented here. Without loss of generality, imagine that two groups, $g = r,f$ where r represents the reference group and f the focal group, exist. Following Chang et al. (2017):

$$\sigma_j^{2(r)} = 1, \quad \text{for } j= 1,..,J \tag{7}$$

$$E(\eta^{(r)}) = 0, \tag{8}$$

$$\lambda_1^{(r)} = \lambda_1^{(f)}, \quad \sigma_1^{2(r)} = \sigma_1^{2(f)}, \quad \tau_{1,c}^{(r)} = \tau_{1,c}^{(f)}, \quad \text{for some } c \in (0,1,2,...,C\text{-}1) \tag{9}$$

$$\sigma_j^{2(r)} = \sigma_j^{2(f)} \text{for } j = 2,..,J. \tag{10}$$

These constraints serve the purpose to set a scale for the latent response variable X_j^*, for the latent construct η and to make the scale comparable across groups. That is, (7) and (8) set the scale of the latent construct η for the reference group, while (9) makes the scale comparable across groups using the anchor. Finally, (10) guarantees a common scale across all the other items. Furthermore, the above-mentioned constraints can be seen as MG-IRT-type constraints where the unique variances σ_j^2 are constrained to be 1 both for the focal and the reference group, the mean of the latent construct η is set to 0 and at least one item is picked as the anchor item, which parameters are set to be equal across groups (Chang et al., 2017).

By means of these constraints the two models are exactly the same. Thus, the remaining differences between MG-CCFA and the MG-noGRM in testing MI can be attributed to the level at which it can be tested (scale vs. item) as well as what testing strategies/measures are used to test it.

2.2 MI hypotheses

Generally, a measurement is said to be invariant if the score that a person obtains on a scale does not depend on his/her belonging to a specific group but only on the underlying psychological construct. Formally, assume that a vector of scores on some items $\mathbf{X}$ is observed, where $\mathbf{X} = \{X_1, X_2,..., X_j\}$, and that a vector of scores on some latent variables $\boldsymbol{\eta}$ underlies these scores, where $\boldsymbol{\eta} = \{\eta_1, \eta_2,...,\eta_r\}$. Then, measurement invariance holds if:

$$P(\mathbf{X}|\boldsymbol{\eta}, \mathbf{g}) = P(\mathbf{X}|\boldsymbol{\eta}). \tag{11}$$

Equation (11) shows that the probability of observing a set of scores $\mathbf{X}$ given the underlying latent construct ($\boldsymbol{\eta}$) is the same across all groups. Moreover, the equation is quite general in the sense that no particular model is yet specified for $P(\mathbf{X}|\boldsymbol{\eta})$.

As discussed above, an equivalent model for $P(\mathbf{X}|\boldsymbol{\eta})$ can be specified for MG-CCFA and the MG-noGRM. Then, one of the main differences in the way these two approaches test MI is whether a test is conducted for the whole vector of scores at once or for each element of the vector separately. The former is more common in MG-CCFA while the latter is generally used within MG-IRT. However, in principle, both types of test can be conducted within each framework.

2.2.1 Scale level. In MG-CCFA MI is tested for all items at once. Different model parameters can be responsible for measurement non-invariance, and they are tested in a step-wise fashion. In each step a new model is estimated, with additional constraints imposed, to test the invariance of a specific set of parameters. The fit of the model to the data is then evaluated to test whether these new constraints worsen it significantly. The latter being true indicates that at least some of the constrained parameters are non-invariant.

2.2.1.1 Configural. The starting point in MG-CCFA is testing configural invariance. In this step the aim is to test whether, across groups, the same number of factors hold and that each factor is measured by the same items. This is generally done by first specifying and then estimating the same model for all groups. Afterwards, fit measures are examined to determine whether the hypothesis of the same model underlying all groups is rejected or not.

2.2.1.2 Metric. If the hypothesis of configural invariance is not rejected, the next step is to test the equivalence of factor loadings. This step is also called the weak or metric invariance step. Commonly, the factor loadings of all items are constrained to be equal across groups. The hypothesis being tested here is that:

$$H_{metric} : \Lambda^{(g)} = \Lambda. \tag{12}$$

If (12) is supported, the equivalence of factor loadings indicates that each measured variable contributes to each latent construct to a similar extent across groups (Putnick & Bornstein, 2016).

2.2.1.3 Scalar. If metric invariance holds, scalar invariance or invariance of the intercepts can be tested. In MG-CCFA, though, the observed data are assumed to come from an underlying continuous response variable X_j^*. This variable does not have a scale and, generally, its intercept is fixed to 0. That is why the thresholds are tested instead of the intercepts. In order to do that, the thresholds for all items are constrained to be equal across groups, while keeping the previous contraints in place. Formally, the hypothesis being tested is:

$$H_{scalar} : T_{j,c}^{(g)} = T_{j,c} \quad \text{for } j = 1,\ 2,\ ..,\ J,\ c = 0,1,2,...,C\text{-}1. \tag{13}$$

If the hypothesis in (13) is not rejected it can be concluded that the thresholds parameters for all items are the same across groups. Finally, it is worth noting that, to obtain full factorial invariance, equivalence of the residual variances should also be tested (Meredith & Teresi, 2006). However, many researchers do not consider this step, since it is not relevant when comparing the mean of the latent constructs across groups (Vandenberg & Lance, 2000).

2.2.2 Item level. In MG-IRT the functioning of each item is tested separately. An item shows differential item functioning (DIF) if the probability of selecting a certain category on that item differs across two groups, given the same score on the latent construct. It is worth noting that, when DIF is tested following a typical MG-IRT-based approach, configural invariance is generally assumed. Also, compared to MG-CCFA where item parameters are firstly allowed to differ and then constrained to be equal across groups, testing DIF follows a different rationale. That is, the starting assumption is that all items function equivalently across groups. Formally:

$$H_0 : \alpha_j^{(g)} = \alpha_j = \frac{\lambda_j^{(g)}}{\sigma_j} = \frac{\lambda_j}{\sigma_j}, \delta_{j,c}^{(g)} = \delta_{j,c} = \frac{-\tau_{j,c}^{(g)}}{\frac{\lambda_j^{(g)}}{\sigma_j}} = \frac{-\tau_{j,c}}{\frac{\lambda_j}{\sigma_j}} \quad (14)$$

for j = 1,2,..,J, c = 0,1,2,...,C-1.

The constraints on one item are then freed up to test whether its parameters are invariant, while keeping the other items constrained to be equal across groups. The procedure is then iteratively repeated for all the other items in the scale. DIF can take two different forms: uniform and nonuniform.

2.2.2.1 Uniform DIF. Given two groups, an ordinal item shows uniform DIF when, between groups, the thresholds parameters differ. In formal terms:

$$H_{no\ uniformDIF} : \delta_{J/k,c}^{(g)} = \delta_{J/k,c} = \frac{-\tau_{J/k,c}^{(g)}}{\frac{\lambda_{J/k}^{(g)}}{\sigma_{J/k}}} = \frac{-\tau_{J/k,c}}{\frac{\lambda_{J/k}}{\sigma_{J/k}}} \quad (15)$$

for j = 1,2,..,J, c = 0,1,2,...,C-1 and for some k, where k = 1,2,...,J.

Where the subscript J/k stands for all items except item k. Equation (15) shows the hypothesis of no uniform DIF indicating that the thresholds of all items except item k ($\tau_{J/k,c}$) are the same across groups. Furthermore, it is interesting to note the connection between uniform DIF and scalar invariance, since both can be seen as tests for shifts in the thresholds parameters.

2.2.2.2 Nonuniform DIF. An ordinal item shows nonuniform DIF when the loading parameter differ across two groups. The tested hypothesis can be formally written as:

$$H_{no\ nonuniformDIF} : \alpha_{J/k}^{(g)} = \alpha_{J/k} = \frac{\lambda_{J/k}^{(g)}}{\sigma_{J/k}} = \frac{\lambda_{J/k}}{\sigma_{J/k}} \quad (16)$$

for j = 1,2,..,J, c = 0,1,2,...,C-1 and for some k, where k = 1,2,...,J.

Equation (16) shows the hypothesis of no nonuniform DIF indicating that for all items except item k the loadings are the same for all groups. This is similar to testing metric invariance in MG-CCFA, but note that here items are evaluated individually.

2.3 MI testing strategies

2.3.1 MG-CCFA-based. Besides commonly testing different hypotheses, MG-CCFA and MG-IRT differ in terms of what testing strategies/measures are used to test these hypotheses. Within MG-CCFA the common strategy is to estimate two nested models and then compare how well they fit the data. A measure of how well a model fits the data is commonly obtained by means of a goodness of fit index. A goodness-of-fit index is a measure of the similarity between the model-implied covariance structure and the covariance structure of the data (Cheung & Rensvold, 2002). To date many fit indices exist, and they can be mainly divided into three categories: measures of absolute fit, misfit and comparative fit (for a more detailed review on the available measures we refer the reader to Schreiber, Nora, Stage, Barlow, & King, 2006).

2.3.1.1 Absolute fit indices. Absolute fit indices focus on the exact fit of the model to the data and one of the most commonly used is the chi-squared (χ^2) test. Imagine that we have a MG-CCFA model A, with χ^2_{ModA} and df_{ModA} indicating the model χ^2 and *degrees of freedom*, which fits sufficiently well the data. To test one of the MI hypotheses (e.g., metric invariance) a new model is specified by constraining the parameters of interest (e.g., loadings) of all items to be equal across groups. Let us call this model B, with χ^2_{ModB} and df_{ModB}. A χ^2 test is then conducted by looking at the difference in these two models:

$$T \sim \chi^2_D(df_D) = \chi^2_{ModB} - \chi^2_{ModA}(df_{ModB} - df_{ModA}). \tag{17}$$

A significant T (e.g., using a significance level of .05) indicates that model B fits significantly worse, and thus that model A should be preferred. This implies that invariance of the constrained parameters (e.g., loadings) does not hold.

Two considerable limitations of the χ^2 test are that, on the one hand, its performance is largely underpowered for small samples because the test statistic is only χ^2-distributed as N goes to infinity (i.e., only with large samples). On the other hand, it is highly strict with large samples indicating, for example, that two models are significantly different even when the differences in the parameters are small.

2.3.1.2 Misfit indices. On top of the well known limitations of the χ^2 test, a general counterargument towards the use of absolute fit indices is that we might not be necessarily interested in the exact fit as much as the extent of misfit in the model (Millsap, 2012). In this case, misfit indices, such as the root mean square error approximation (RMSEA) can be used. This index quantifies the misfit per degrees of freedom in the model (Browne & Cudeck, 1993). Specifically, in the case of multiple groups, it can be expressed as:

$$RMSEA = \sqrt{G}\ \sqrt{max\left[\frac{\chi^2_{ModA}}{df_{ModA}} - \frac{1}{N-1}, 0\right]}. \tag{18}$$

Based on which MI hypothesis is tested, different criteria and procedures are used to determine whether the RMSEA is acceptable. In the configural step, the absolute value of RMSEA is used. Specifically, values between 0 and .05 indicate a “good" fit and values between .05 and .08 are thought to be a “fair" fit (Browne & Cudeck, 1993; Brown, 2014). In the subsequent steps, the change in the RMSEA (ΔRMSEA) is used instead of the absolute value of the measure. Specifically, a ΔRMSEA of .01 has been suggested as a cut-off value in the case of metric invariance and, similarly, a value of .01 should be used for scalar invariance (Cheung & Rensvold, 2002 Chen, 2007). When the change in the ΔRMSEA is higher than the specific cut-off, invariance is rejected.

2.3.1.3 Comparative fit indices. The third category of fit indices is the one of comparative fit, where the improvement of the hypothesized model compared to the null model is used as an index to test MI. Differently from exact fit indices, where the hypothesized model is compared against a saturated model (a model with $df = 0$), in comparative fit indices a comparison is conducted between the hypothesized model and the null model, with $\chi^2_{ModNull}$ and $df_{ModNull}$. The null model is a model in which all the measured variables are uncorrelated (i.e., a model where there is no common factor). Numerous of these measures exist and, among them, a well-known one is the comparative fit index (CFI) (Bentler, 1990). The CFI measures the overall improvement in the χ^2 in the tested model compared to the null model, and can be formally written as:

$$CFI = 1 - \frac{\chi^2_{ModA} - df_{ModA}}{\chi^2_{Null} - df_{Null}} \tag{19}$$

where a value of .95 is used as a cut-off value in the configural invariance step to indicate a “good" fit (Bentler, 1990). In the subsequent steps, the common guidelines for cut-off values focus on the change in CFI (ΔCFI). Specifically, a ΔCFI higher than .01 is considered to be problematic both in the case of testing for loadings and thresholds invariance (Cheung & Rensvold, 2002 Chen, 2007).

2.3.2 MG-IRT-based. In MG-IRT-based approaches both parametric and nonparametric methods exist to test for uniform and nonuniform DIF. In this paper the focus is on parametric methods, where a statistical model is assumed. Specifically, methods that compare the models' likelihood functions will be discussed (for a more detailed discussion on both parametric and nonparametric methods for DIF detection, we refer the reader to Millsap, 2012).

2.3.2.1 Likelihood-Ratio test. One well known technique for the study of DIF is the likelihood-ratio test (LRT) (Thissen, Steinberg, and Gerrard 1986; Thissen 1988; Thissen, Steinberg, and Wainer 1993). In this test, the log-likelihood of a model with the parameters of all items constrained to be equal across groups is compared against the log-likelihood of the same model with freed parameters for one item only. The former is sometimes called the compact model (L_C), while the latter is sometimes called the augmented model (L_A, S.-H. Kim and Cohen 1998; Finch 2005). Once these two models are estimated and the log-likelihood (lnL_C and lnL_A) is obtained, the test statistic (G^2) can be calculate using the following formula:

$$G^2 = -2lnL_C - (-2lnL_A) = -2lnL_C + 2lnL_A. \tag{20}$$

The G^2 is χ^2 distributed with *df* equal to the difference in the number of parameters estimated in the two models (Thissen, 1988). The same procedure is then iteratively repeated for all items. It is important to highlight that the above equation represents a omnibus test of DIF, which in case of a significant result could be further inspected by constraining only specific parameters. For example, it would be possible to test uniform DIF by allowing only the thresholds to vary across groups.

2.3.2.2 Logistic regression. Logistic regression (LoR; Swaminathan & Rogers, 1990) is another parametric approach that has recently gained interest among DIF experts (Yasemin, Leite, & Miller, 2015). The intuition behind the LoR approach is similar to the one of step-wise regression in which one can test whether the model improves by sequentially entering new predictors. The common order in which the variables are introduced, starting with a null model where only the intercept is estimated, is by first adding the latent construct, then the grouping variable, and finally an interaction between the latent construct and the grouping variable. Formally, this sequence of models is written as:

$$Model\ 0 : logitP(y_j \geq c) = \nu_c; \tag{21}$$

$$Model\ 1 : logitP(y_j \geq c) = \nu_c + \beta_1\eta; \tag{22}$$

$$Model\ 2 : logitP(y_j \geq c) = \nu_c + \beta_1\eta + \beta_2 G; \tag{23}$$

$$Model\ 3 : logitP(y_j \geq c) = \nu_c + \beta_1\eta + \beta_2 G + \beta_3\eta G. \tag{24}$$

In the equations above $P(y_j \geq c)$ is the probability of the score on item j falling in category c or higher, and ν_c is a category specific intercept. It is worth to point out that, compared to the LRT, the latent variable scores are in this case only estimated once and then treated as observed. One clear disadvantage is that, since the latent variable scores are estimated and not observed, there might be uncertainty in the estimates, which could, in turn, affect the performance of this method. Moreover, some alternative formulations make use of sum scores instead of estimates of latent variable scores (Rogers & Swaminathan, 1993). Once the logistic regression models are estimated and a G^2 is obtained, an omnibus DIF test can be conducted by:

$$G^2_{omnibus} = G^2_{Model3} - G^2_{Model1}, \tag{25}$$

which is asymptotically χ^2 distributed with df=2 (Swaminathan & Rogers, 1990). Zumbo (1999) suggested to investigate the source of bias by separately testing for uniform and nonuniform DIF, respectively:

$$G^2_{uniDIF} = G^2_{Model2} - G^2_{Model1} \tag{26}$$

and:

$$G^2_{nonuniDIF} = G^2_{Model3} - G^2_{Model2} \tag{27}$$

where both (26) and (27) are χ^2 distributed with df=1.

The omnibus test procedure (25) turned out to have an inflated number of incorrectly flagged DIF items (Type I error; Li and Stout 1996). To solve this issue, a combination of a significant 2-df LRT (25) and a measure of the magnitude of DIF using a pseudo-R^2 statistic has been suggested as an alternative criterion (Zumbo, 1999). The underlying idea is to treat the β coefficients as weighted least squares estimates and look at the differences in pseudo-R^2 (ΔR^2) measures between the model with and without the added predictor (e.g., Cox & Snell, 1989). Specifically, to flag an item as DIF, both a significant χ^2 test (with df=2) and an effect size measure with an ΔR^2 of at least .13 is suggested to be used (Zumbo, 1999).

3 Simulation studies

Two simulation studies were performed to evaluate the impact of MG-CCFA- and MG-IRT-based hypotheses and testing strategies on the power to detect violations of MI. In the first study, an invariance scenario was simulated where parameters were invariant between groups. In the second study, a non-invariance scenario was simulated where model parameters varied between groups.

3.1 Simulation Study 1: invariance

In the first study three main factors were manipulated:

1. The number of items at 2 levels: 5, 25, to simulate a short and a long scale;

2. The number of categories for each item at 2 levels: 3, 5;

3. The number of subjects within each group at 2 levels: 250, 1000.

A full-factorial design was used with 2 (number of items) x 2 (number of categories) x 2 (number of subjects within each group) = 8 conditions. For each condition 500 replications were generated.

3.1.1 Method.

3.1.1.1 Data generation.

Data were generated from a factor model with one factor and two groups. The population values of the model parameters were chosen prior to conducting the simulation study and are reported in Table 1. The choice of the values began with specifying the standardized loadings. Specifically, they were selected to resemble the ones commonly found in real applications with items having medium to high correlation with the common factor but differing among them.

The second step was to select the thresholds and, in order to choose them, continuous data with 10,000 observations were firstly generated under a factor model using the loadings in Table 1. Afterwards, using the distribution of the item scores for item 1, which was subsequently used as the anchor item, the tertiles (for items with three categories) and the quintiles (for items with five categories) were calculated. In particular, the generation of the remaining thresholds proceeded by shifting the tertiles/quintiles of the first item by half a standard deviation. In detail, for both the three- and five-categories case, we shifted the thresholds value of the second and fifth item by + .50 and of the third and fourth item by - .50 (as can be seen from Table 1). In the conditions with 25 items, the same parameters in Table 1 were repeated five times. In all estimated models item 1 was used as the anchor item.

3.1.1.2 Data analysis.

Scale level. 3.1.1.2.1 The specification of the MG-CCFA models to test MI followed the common steps of a general MI testing procedure as described in Section 2.2.1. Specifically, in the configural step, a unidimensional factor model was fitted to both groups allowing loadings and thresholds to differ between groups (configural invariant model).

In the metric step, factor loadings were constrained to be equal across groups while allowing the thresholds to be freely estimated (metric invariant model). In the scalar step, both factor loadings and thresholds were constrained to be equal across groups (scalar invariant model). Afterwards, a χ^2 test ($\alpha = .05$) was conducted between: (i) the model estimated in the configural and the metric step to test for loadings invariance, and (ii) the model estimated in the metric and scalar scalar step to test for thresholds invariance. Additionally, the change in RMSEA (ΔRMSEA) and in CFI (ΔCFI) was calculated between the just mentioned models. Loadings non-invariance was concluded if at least one of the following criteria was met: a significant χ^2 test, a $\Delta\text{RMSEA} > .01$ or a $\Delta\text{CFI} > .01$. Additionally, since the common guidelines reported in the literature recommend to base decisions about (non)invariance of parameters using various indices, a combined criterion was created. According to this combined criterion, loadings non-invariance at the scale level was concluded if both a significant χ^2 test and at least one between a $\Delta\text{RMSEA} > .01$ or a $\Delta\text{CFI} > .01$ was found (Putnick & Bornstein, 2016). Thresholds non-invariance at the scale level was concluded if at least one of the following criteria was met: a significant χ^2 test, a $\Delta\text{RMSEA} > .01$ or a $\Delta\text{CFI} > .01$. Also, in this case a combined criterion was created. Specifically, a scale was considered non-invariant with respect to thresholds if both a significant χ^2 and at least one between a $\Delta\text{RMSEA} > .01$ or a $\Delta\text{CFI} > .01$ was found. In addition, all MG-CCFA models were estimated using diagonally weighted least square (DWLS). This is a two-step procedure where in the first step the thresholds and polychoric correlation matrix are estimated and then, in the second step, the remaining parameters are estimated using the polychoric correlation matrix from the previous step.

In MG-IRT-based procedures MI is tested for each item individually. Therefore, to conduct a test at the scale level, we decided to flag the scale as non-invariant if at least one item was flagged as non-invariant, correcting for multiple testing. Two different testing strategies were considered: the logistic regression (LoR) procedure and the likelihood-ratio test (LRT). Within LoR, two different criteria were used to flag an item as non-invariant. The first criterion is based on the likelihood-ratio test (LRT). Specifically, an item was non-invariant, either with respect to loadings or thresholds, in the case of a sig-

nificant χ^2 test ($\alpha = .05$) between a model where the latent construct score, the grouping variable and an interaction between the two are included (24) and a model with only the latent construct score (22) (Swaminathan & Rogers, 1990). The second criterion, which will from this point on be called R^2, combines the just mentioned χ^2 test with a measure of the magnitude of DIF. The latter is obtained by computing the difference between a pseudo-R^2 measure between the two above mentioned models (ΔR^2). Using this approach, an item was flagged as non-invariant when both a significant χ^2 test and a $\Delta R^2 > .02$ were found (Choi, Gibbons, & Crane, 2011). Specifically, in this simulation study, the McFadden pseudo-R^2 measure was used (Menard, 2000). In the case of the LRT, two different models per item were estimated. In one model the constraints on the thresholds were released for a specific item (uniform DIF model), while in the other the constraint on the loading was released (nonuniform DIF model). Additionally, a model with all items constrained to be equal was estimated (fully constrained model). An item was flagged as non-invariant with respect to thresholds in case of a significant LRT ($\alpha = .05$) between the fully constrained model and the uniform DIF model. Similarly, an item was flagged as non-invariant with respect to loadings in case of a significant LRT ($\alpha = .05$) between the fully constrained model and the nonuniform DIF model. This procedure was repeated iteratively for all the other items. Since multiple tests are conducted for the scale, a Bonferroni correction was used.

Item level. 3.1.1.2.2 In order to test MI at the item level using a MG-CCFA-based testing strategy a backward/step-down procedure was used (E. S. Kim & Yoon, 2011; Brown, 2014). The rationale is the same as the one just described in the LRT for MG-IRT. Specifically, the constraints (either on the thresholds or on the loading) were released for only one item, while keeping all the other items constrained to be equal. Hence, for each item two different models were estimated. Then, the χ^2 test ($\alpha = .05$) was conducted and the ΔRMSEA and ΔCFI calculated. This procedure was then repeated iteratively for all the other items. Note that, due to the multiple tests conducted, Bonferroni correction was used. For MG-IRT-based procedures, the same procedures and criteria used at the scale level were used to test MI at the item level (but without applying a Bonferroni

correction).

3.1.1.3 Outcome measures. The convergence rate (CR) and the false positive rate (FPR) were calculated both for MG-CCFA- and MG-IRT-based procedures both at the scale and at the item level. The CR indicates the proportion of models that converged while the FPR represents the scales/items incorrectly flagged as non-invariant. If models did not converge, new data were generated and models were rerun in order to always calculate the FPR based on 500 repetitions.

3.1.1.4 Data simulation, softwares and packages. The data were simulated and analyzed using R (R Core Team, 2013). Specifically, for estimating MG-CCFA and obtaining fit measures the R package *lavaan* was used (Rosseel, 2012), while for LoR and the LRT *lordif* (Choi et al., 2011) and *mirt* (Chalmers, 2012) were used, respectively.

3.1.2 Results.

3.1.2.1 Convergence Rate. The convergence rate was almost 100% for all the considered approaches across all the conditions. Models' non-convergence was observed only for a few conditions with small sample size as well as short scales and never exceeded 1%. The tables showing the complete results can be found in the appendix (Tables A1 through A4)

3.1.2.2 Scale level performance. The scale-level results when loadings equivalence was tested are reported in Table 2. For MG-CCFA, with long scales only the χ^2 test produced a high number of false positives. Hence, long scales were falsely flagged as non-invariant while no differences existed in the population parameters. Also, ΔRMSEA and ΔCFI showed a FPR higher than the chosen α level for small sample size and short scales. Within MG-IRT-based approaches, the results were quite different, depending on the testing strategy. For the LoR approach, using the LRT criterion, the results obtained in this simulation study aligns with the ones in the existing literature, with an evident inflation of the FPR (overall, FPR $>$ .40) (Rogers & Swaminathan, 1993; Li & Stout, 1996). For the R^2 criterion, where a combination of the LRT and a pseudo-R^2 measure was used, the FPR was at or below the chosen α level using the R^2 criterion, with an inflated FPR only in the case with $N = 250$, $C = 3$ and $J = 25$ (FPR = 0.182). One

possible explanation is that, due to the small amount of information available for each person in this condition there is more uncertainty in the estimated scores of the latent construct. Since these estimates are then used as observed variables in the LoR procedure, they are likely to produce a larger number of items incorrectly flagged as non-invariant. Finally, the LRT showed an acceptable FPR in all conditions when testing for loadings equivalence at the scale level.

The results of the simulation study when equivalence of thresholds was tested at the scale level are reported in Table 3. For MG-CCFA, the FPR was always above .10 level using the χ^2 test with long scales compared to short scales. Also, ΔRMSEA showed a FPR higher than the chosen α level with short scales, while the ΔCFI showed a FPR higher than the chosen α level in the conditions with both short scales and small sample size. The combined criterion seemed to be the only one, across the ones used for MG-CCFA, that provided an acceptable FPR rate across conditions with the only exception when $N = 1000$, $C = 5$ and $J = 25$ (FPR = .102). For MG-IRT-based testing strategies, the obtained results are similar to the ones observed in the case of testing loadings equivalence. Specifically, for the LoR approach, the R^2 criterion performed well in all conditions except when $N = 1000$, $C = 3$ and $J = 5$ (FPR = .189). Moreover, the LRT criterion for LoR showed an evident inflation across all conditions. Finally, the LRT performed well in all conditions.

3.1.2.3 Item-level performance. The results when loadings equivalence was tested at the item level are reported in Table 4. Similarly to the results observed at the scale level, for MG-CCFA the FPR appeared to be highly inflated using the χ^2 test with long scales. For MG-IRT using the LoR procedure, the LRT criterion produced a high number of false positives with short scales. Moreover, the results for both the R^2 criterion and the LRT were within the chosen α level in all conditions.

Finally, the results when testing thresholds equivalence at the item level are reported in Table 5. For MG-CCFA, all the criteria performed reasonably well with some small inflations of the χ^2 test. A FPR slightly higher than the chosen α level was also observed for ΔRMSEA and ΔCFI in the condition with both short scales and small sample. For

MG-IRT-based testing strategies, only the LRT criterion for the LoR approach showed a FPR higher than the chosen α level with $J = 5$.

3.2 Simulation Study 2: non-invariance

In the second simulation study, three more factors were included to evaluate the performance of the studied approaches, with their respective testing strategies, in detecting violations of MI when parameters were non-invariant across groups. On top of varying the scale length, the number of categories and the sample size we now also vary:

1. Percentage of items with non-invariant loadings at 3 levels: 20%, 40% aligned, and 40% misaligned;

2. Percentage of items with non-invariant thresholds at 3 levels: 20%, 40% aligned, and 40% misaligned;

3. The amount of bias imposed for each non-invariant parameter at two levels: small and large.

The first three factors were the ones used in the previous simulation study. Additionally, to simulate differences in loadings/thresholds across groups the values of the parameters were changed either for 20% or 40% of the items. Moreover, in the condition with 40% of the items having non-invariant loadings, the values were either increased for all items (e.g., all loadings on one group are higher), or increased for half of the items and decreased for the other half (e.g., in the condition with 5 items, where the values of two loadings are changed, one was increased and the other decreased). The former was labeled as an aligned change while the latter as a misaligned change.

The same procedure was followed for the shifts in thresholds both in terms of percentage of items with non-invariant thresholds and for the aligned or misaligned shifts. Note that, since each item has more than one threshold, all the thresholds of that item were shifted. The manipulated violations of MI, both for loadings and thresholds, were either small or large. On the one hand, a difference of .1 or .2 was used to simulate small and large changes in the standardized factor loadings, respectively. The chosen values substantially

increase the variance accounted by the factor for the item. For example, in a standardized factor loading of .7 the explained variance of the item by the factor is $.7^2 = .49$. If the loading is increased by .1 the explained variance will then be $.8^2 = .64$. Also, in case of a big change (.2), the explained variance will become $.9^2 = .81$. On the other hand, for the shifts in thresholds, the parameters of one group were shifted by either a quarter (.25) or half a standard deviation (.50) to simulate small and large violations of thresholds non-invariance.

In total, 2 (number of items) x 2 (number of categories) x 2 (number of subjects within each group) x 3 (percentage of non-invariant loadings) x 3 (percentage of non-invariant thresholds) x 2 (amount of bias imposed) = 144 conditions were simulated for the conditions with non-invariance in the loadings and the thresholds. For each condition 500 replications were generated.

3.2.1 Method.

3.2.1.1 Data analysis. Like in the first simulation study, the data were generated from a factor model with one factor and two groups. The population parameters were the same as used in the first simulation study and they were varied, based on the condition, as just explained above. Moreover, the procedures used to specify and estimate the models, both at the scale and at the item level, were the same ones used previously. Differently from before, only a subset of the criteria was used to flag a scale/item as non-invariant. Specifically, only the criteria that showed an acceptable FPR across all conditions in the first simulation study are reported. This was done because procedures with unacceptable FPRs should not be considered for testing MI, and hence considering them here would not make sense. Thus, for MG-CCFA only the results of the combined criterion are reported, while for MG-IRT-based procedures the LRT approach and, for the LoR approach, only the results of the R^2 criterion.

3.2.1.2 Outcome measures. The convergence rate (CR), true positive rate (TPR) and false positive rate (FPR) were calculated both for the MG-CCFA- and MG-IRT-based procedures both at the scale and at the item level. Here, the TPR represents the proportion of non-invariant scales/items that are correctly identified as such, while the

FPR represents the proportion of non-invariant scales/items that are incorrectly identified as such. If models did not converge, new data were generated and models were rerun in order to always calculate the TPR and the FPR for 500 repetitions.

3.2.2 Results.

3.2.2.1 Convergence Rate.

Scale level. 3.2.2.1.1 The results of the CR when testing loadings equivalence at the scale level in the non-invariance scenario are displayed in Table A5 in the Appendix. In the conditions with large sample size, the CR when testing loadings equivalence at the scale level was 99% for all the approaches. Compared to the conditions with a large sample size, the CR dropped in the conditions with small sample size and 40% of the items showing large misaligned changes in loadings. Specifically, the CR for MG-CCFA was .978 when $J = 5$ and $C = 3$ while for MG-IRT using the LoR approach the CR was around .9 with $N = 250$, $J = 25$ and both for items that had 3 or 5 categories.

The results of the CR when testing thresholds equivalence at the scale level in the non-invariance scenario are displayed in Table A6 in the Appendix. For MG-CCFA, the CR was generally lower in the conditions with large shifts in the thresholds compared to the conditions with small shifts. For example, with $N = 250$, $J = 5$, and large shifts in the thresholds parameters the CR was .808. This lower CR could be due to a specific issue with the estimation procedure. In fact, using DWLS, the estimation heavily relies on the first step, where the the thresholds and the polychoric correlation matrix are estimated. Large differences in thresholds between the two groups might affect this first step and, in turn, the remaining part of the procedure. On the contrary, for MG-IRT-based approaches the CR was always above 99%.

Item level. 3.2.2.1.2 The results of the CR when testing loadings equivalence at the item level in the non-invariance scenario are displayed in Table A7 in the Appendix. These results closely resemble the ones observed when loadings equivalence were tested at the scale level. Specifically, the CR was below .98 for MG-CCFA only in the condition with $N = 250$, $C = 3$, $J = 5$, and large misaligned changes in loadings in 40% of the items. Moreover, for MG-IRT using the LoR approach the CR was around .89 when $N =$

250, $J = 25$, and with large misaligned changes in the loadings, regardless of the number of categories for each item.

The results of the CR when testing thresholds equivalence at the item level in the non-invariance scenario are displayed in Table A7 in the Appendix. For MG-CCFA, similar to what was observed at the scale level, the CR dropped in the conditions with small sample size, big shifts in thresholds and short scales compared to the other conditions. For example, the lowest CR was observed in the condition with $N = 250$, $C = 3$, $J = 5$ and large misaligned shifts in thresholds (CR = 0.796). However, for MG-IRT-based approaches the CR was always above 99%.

3.2.2.2 Scale-level performance. The results of the TPR when testing loadings equivalence at the scale level in the non-invariance scenario are displayed in Table 6. Overall, in the conditions with small changes in the loadings, the TPR for all the considered approaches was below .4. For example, the highest TPR was observed using the LRT, which was only .398. Since, in the non-invariance scenario, for MG-CCFA, a combined criterion was used to flag scales/items as non-invariant, we further inspected the TPRs for each of the measures that form this combined criterion. These results are displayed in the appendix in Table A11. For ΔCFI, the results seemed to highly depend on the length of a scale, especially when aligned changes were simulated in 20% and 40% of the loadings. For short scales, MG-CCFA generally outperformed both MG-IRT approaches, regardless of the other factors. For long scales, MG-IRT LRT outperformed both MG-CCFA and MG-IRT LoR. Also, since in the first simulation study the LoR approach with $N = 250$, $J = 5$ and $C = 3$ had an unacceptable FPR, the results in this simulation study are reported in red indicating that they should not be considered.

The results of the TPR when testing thresholds equivalence at the scale level in the non-invariance scenario are displayed in Table 7. MG-CCFA outperformed both MG-IRT approaches in the conditions with short scales, regardless of the other factors. However, for long scales the LRT for MG-IRT outperformed both MG-CCFA and the LoR approach. In addition, LoR's TPR was lower than the one of MG-CCFA and the LRT, in almost all conditions, and especially when the sample size was big. However, in the case of

misaligned shifts the TPR was almost always the same as it was for MG-CCFA and the LRT.

3.2.2.3 Item-level performance. The results of the TPR when testing loadings equivalence at the item level in the non-invariance scenario are displayed in Table 8. The results of the FPR were also calculated and are displayed in Table A9 in the Appendix. Similar to the results observed at the scale level, MG-CCFA hardly detects non-invariance whensmall changes in the loading exist at the item level, reaching a maximum TPR of .479 with misaligned changes affecting 40% of the items, $N = 1000$, $C = 5$ and $J = 5$. Furthermore, difficulties in flagging non-invariant items were even more pronounced for long scales, showing that loadings nonequivalence was not detected in most cases. Similar to what was done at the scale level, the performance of ΔRMSEA and ΔCFI, for MG-CCFA, was further investigated. These results are displayed in the appendix in Table A12. For both fit measures, the results seemed to highly depend on the length of a scale. In fact, for long scales, both measures rarely detected changes in loadings. For MG-IRT-based approaches, differences in loadings were rarely detected by the LoR approach regardless of the condition, and with even lower frequencies when the sample size increases. The LRT outperformed both MG-CCFA and LoR in all conditions in terms of the TPR.

The results of the TPR when testing thresholds equivalence at the item level in the non-invariance scenario are displayed in Table 9. The results of the FPR were also calculated and are displayed in Table A10 in the Appendix. For short scales MG-CCFA outperformed both LoR and LRT regardless of the other factors. However, for longer scales, the LRT had a higher TPR compared to the LoR approach and MG-CCFA.

3.3 Conclusion

Based on the results observed in the invariance scenario, we can conclude that, for only some of the MG-CCFA- and MG-IRT-based testing strategies a FPR below or at the chosen α level was found. In fact, among the considered testing strategies used to flag a scale/item as non-invariant, quite many methods had an inflated type I error. For

MG-CCFA-based criteria, the FPR was often below or at the chosen α level when a combination of a χ^2 test and an alternative fit measure (e.g., RMSEA or CFI) was used. Only in one case, with large sample size ($N = 1000$), a short scale (5 items) and items with five categories an inflation of the FPR was noted (FPR = .102). Additionally, using only the χ^2 test with long scales (25 items) produced a high number of false positives, especially when loadings non-invariance was tested. For MG-IRT, the LRT provided a well-controlled FPR in all conditions regardless of whether the test was conducted at scale or at the item level. The LoR approach for MG-IRT showed an inflated FPR when the LRT criterion was used, while adopting a combination of both the LRT criterion and a pseudo-R^2 measure resulted in a low FPR in (almost) all conditions.

Based on the results observed in the non-invariance scenario, we can conclude that, when testing loadings equivalence, all the studied approaches hardly detect small changes in loadings. Furthermore, when loadings equivalence was tested at the scale level, MG-CCFA outperformed both MG-IRT-based approaches for short scales, while, for long scales, MG-IRT LRT outperformed MG-CCFA and MG-IRT LoR. Additionally, when loadings equivalence was tested at the item level, MG-IRT LRT generally outperformed the other approaches. Finally, when thresholds equivalence was tested, MG-CCFA outperformed both MG-IRT-based approaches for short scales, while, for long scales, MG-IRT LRT outperformed MG-CCFA and MG-IRT LoR.

The results of the simulation studied were summarized in a flowchart (Figure 1) and its paths will be now discussed to facilitate the reader's interpretation. As indicated in the second node of the flowchart, configural invariance is tested only with MG-CCFA. In fact, for MG-IRT-based testing strategies, configural invariance is commonly assumed.

If configural invariance is not tested, one of the most relevant factors in deciding what approach could be preferred, according to our results, is the length of a scale. As indicated by the right branch of the third node, for long scales, MG-IRT LRT outperformed both MG-CCFA and MG-IRT LoR. Since, in the non-invariance scenario, for MG-CCFA, a combined criterion was used to flag scales/items as non-invariant, we further inspected the TPRs for each of the measures that form this combined criterion. These results, for

the scale- and item-level tests, are displayed in the appendix in Table A11 and Table A12, respectively. In particular, the TPRs for ΔRMSEA and ΔCFI were heavily affected by both scale length and the level at which loadings equivalence was tested (scale or item). Specifically, for long scales, these two measures hardly detected changes in loadings, especially when the test was conducted at the item level. Thus, we would discourage researchers the use of these fit measures, in particular when testing MI for each item individually. In addition, in our simulation studies, the length of the scale was varied only at two levels (5,25). For this reason, we advise the reader to be cautious in generalizing these results to scales of different lengths.

For short scales, choosing a specific approach should depend on which parameters are tested for MI (i.e., loadings or thresholds). The results, in terms of the TPR, specifying which approach performed better and for which parameters, are indicated by the three branches of the fourth node. If loadings are tested, a different approach should be preferred based on the level at which MI is tested (i.e., scale or item). When the test was conducted at the scale level, MG-CCFA outperformed both MG-IRT-based approaches, while, if the test was conducted at the item level, MG-IRT LRT outperformed both MG-CCFA and MG-IRT LoR. It should be pointed out that, the simulated differences in loadings were, overall, small in magnitude (e.g., at the largest .2). It should not come as a surprise, then, that overall the TPR for all the approaches was quite low (FPR $<$ .40), especially in the case of small changes. When thresholds equivalence was tested for short scales, MG-CCFA outperformed both MG-IRT-based approaches. This result is summarized by the right branch of the fourth node.

Finally, if both loadings and thresholds are tested, for short scales, we recommend to use MG-CCFA. This is specified by the central branch of the fourth node. In fact, in terms of detecting loading differences, MG-CCFA outperformed both MG-IRT-based approaches when the test was conducted at the scale level, while, at the item level, it performed slightly worse than (or at least as good as) MG-IRT-based approaches. In addition, for thresholds differences, MG-CCFA (almost) always outperformed both MG-IRT-based approaches. Importantly, the combination of non-invariant loadings and thresholds was

not explicitly tested in the simulation studies. Still, in common practice, both could be present and are generally tested using one specific approach. For these reasons, it is useful to indicate what might be, based on the observed results, the preferred approach to test for both loading and threshold differences.

4 Discussion

When comparing psychological constructs across groups, testing for measurement invariance (MI) plays a crucial role. With ordinal data, multiple group categorical confirmatory factor analysis (MG-CCFA) and multiple group item response theory (MG-IRT) models can be made equivalent using a set of minimal identification constraints (Chang et al., 2017). Still, differences between these two approaches exist in the context of MI testing. These differences are reflected in: (i) the hypotheses being tested, and (ii) the testing strategies/measures used to test these hypotheses. In this paper, two simulation studies were conducted to evaluate the performance of the different testing strategies and measures in testing MI when: (i) the test is conducted at the scale or at the item level and, (ii) MG-CCFA- or MG-IRT-based testing strategies are used. In the first simulation study, an invariance scenario was simulated where no differences existed in the parameters across groups. In addition, a second simulation study was conducted to assess the performance of these approaches when non-invariance was simulated between groups.

A key result of these simulation studies, is that the performance of MG-CCFA- and MG-IRT-based testing strategies and measures mostly depends on the length of a scale. In fact, the likelihood ratio test (LRT) procedure for MG-IRT outperformed both the logistic regression (LoR) procedure and MG-CCFA for long scales, while, for short scales, the results differed based on the parameters being tested (i.e., loadings or thresholds).In general we recommend, based on the observed results, to use MG-CCFA for short scales.

In addition, another key result pertains to how the length of a scale and the level at which MI is tested affects the performance of MG-CCFA's fit measures. In fact, both RMSEA and CFI hardly detected non-invariant parameters when MI was tested for each item individually, especially with long scales. That is, the more items on a scale, the

harder it is, for these measures, to detect whether a specific item is non-invariant. These results identify a fundamental issue when using these fit measures to test MI at the item level. In fact, the cut-off values that are commonly used seem to be inadequate for item-level testing, since their performance heavily depends on the scale's length. Commonly, MG-CCFA is used to test for MI at the scale level, which might explain why most papers focused on defining optimal cut-off values for these measures when MI is tested at this level (Cheung & Rensvold, 2002; Chen, 2007; Rutkowski & Svetina, 2014; Rutkowski & Svetina, 2017). If non-invariance is detected, researchers might decide to inspect its source by conducting a test for each item individually (E. S. Kim & Yoon, 2011; Putnick & Bornstein, 2016). Based on our results, we would discourage researchers from using such measures to this aim since the cut-off values need to be re-evaluated for item-level testing in future research.

The simulation studies conducted provide a useful indication in terms of the performance of testing strategies and measures in testing MI for model applied to ordinal data. Still, they are not free of limitations and it is relevant to highlight some of those. For example, we focused on unidimensional scales, while researchers are frequently confronted with scales that capture multiple dimensions. Generally, MG-CCFA is used for multidimensional constructs, while MG-IRT-based models are preferred with unidimensional constructs. It might therefore be interesting to inspect if similar results as the ones observed here would be obtained when model complexity is increased by having multiple dimensions.

Another set of limitations pertains to the grouping. Firstly, in the current simulation studies we inspected the performance of MG-CCFA- and MG-IRT-based testing strategies with only two groups. However, cross-cultural and cross-national data, where many groups are compared simultaneously, are rapidly increasing in psychological sciences. For this reason, it might be useful to investigate differences in the performance of the studied approaches when many groups are compared. Secondly, in these simulation studies we knew which subject belonged to which group, and differences were created between the groups' measurement models. However, the grouping of subjects is not always known

and/or researchers might not have access to those variables that are thought to cause heterogeneity (e.g., nationality, gender). In this case a different approach might be preferred to disentangle the heterogeneity across participants (e.g., factor mixture models; Lubke & Muthén, 2005).

One last important set of limitations concern the anchoring of the scale. That is, which items' parameters are set equal across groups in order to identify the model and to make the scale comparable across groups. First, the item that was used as the anchor in the simulation studies was known to be invariant across groups. In real applications this information is never known beforehand, and estimating a model relying on an inadequate anchor item could impact model's convergence as well as the ability to detect non-invariance of parameters. This issue has been partly discussed in previous studies comparing different type of identification constraints (Chang et al., 2017). It could be interesting to inspect how the choice of a "good" or "bad" anchor item influences the detection of MI in a more comprehensive study. Second, in these simulation studies, a set of minimal constraints was used to make the measurement models equivalent, and only one item was constrained to be equal across groups. Minimal constraints allow most parameters to be freely estimated. However, when specific items are known to function similarly across groups (e.g., knowledge based on prior studies or strong motivations to consider them invariant across groups) it might be beneficial, both in terms of the estimation and the power to detect non-invariance of the model's parameters, to constrain them to be equal across groups. To our knowledge, the choice of what item(s) should be constrained is often neglected in MG-CCFA where, commonly, the first item is picked without specific theoretical or statistical knowledge of its invariance. In MG-IRT more attention is devoted to this topic. Still, further research might be dedicated to investigate thoroughly how the choice of the anchor (e.g., how many items? what happens if the chosen ones are non-invariant?) affects the performance of both frameworks in detecting MI.

Open practices: The code and data can be made available upon request.

5 References

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

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), S176–S181.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, *154*, 136–136.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chang, Y.-W., Hsu, N.-J., & Tsai, R.-C. (2017). Unifying differential item functioning in factor analysis for categorical data under a discretization of a normal variant. *Psychometrika*, *82*(2), 382–406.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An r package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of Statistical Software*, *39*(8), 1.

Cox, D. R., & Snell, E. J. (1989). Analysis of binary data (vol. 32). *Monographs on Statistics and Applied Probability*.

Finch, H. (2005). The mimic model as a method for detecting dif: Comparison with mantel-haenszel, sibtest, and the irt likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary*

Journal, *15*(1), 136–153.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical cfa and irt. *Structural Equation Modeling*, *18*(2), 212–228.

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345–355.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing dif. *Psychometrika*, *61*(4), 647–677.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21.

Meade, A. W., & Lautenschlager, G. J. (2004). Same question, different answers: Cfa and two irt approaches to measurement invariance. In *19th annual conference of the society for industrial and organizational psychology* (Vol. 1).

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*(1), 17–24.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, S69–S77.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and mantel-haenszel procedures for detecting differential item functioning. *Applied Psycholog-*

ical Measurement, *17*(2), 105–116.

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling and more. version 0.5–12 (beta). *Journal of Statistical Software*, *48*(2), 1–36.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*, *30*(1), 39–51.

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

San Martín, E., & Rolin, J.-M. (2013). Identification of parametric rasch-type models. *Journal of Statistical Planning and Inference*, *143*(1), 116–130.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323–338.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361–370.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

Thissen, D. (1988). Use of item response theory in the study of group differences in trace lines. *Test Validity*.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational

research. *Organizational Research Methods*, *3*(1), 4–70.

Yasemin, K., Leite, W. L., & Miller, M. D. (2015). A comparison of logistic regression models for dif detection in polytomous items: the effect of small sample sizes and non-normality of ability distributions. *International Journal of Assessment Tools in Education*, *2*(1).

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (dif): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. *Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense*.

Table 1

Population values

Item	**3 categories**			**5 categories**			
	λ	τ_1	τ_2	τ_1	τ_2	τ_3	τ_4
1	.5	-0.38	0.38	-0.84	-0.25	0.25	0.84
2	.7	0.12	0.88	-0.34	0.25	0.75	1.34
3	.6	-0.88	-0.12	-1.34	-0.75	-0.25	0.34
4	.4	-0.88	-0.12	-1.34	-0.75	-0.25	0.34
5	.3	0.12	0.88	-0.34	0.25	0.75	1.34

Table 2

Loadings' FPR scale level - invariance scenario

			FPR scale level- loadings						
				MG-CCFA			**MG-IRT LoR**		**MG-IRT LRT**
N	**C**	**J**	**Comb**	χ^2	**ΔRMSEA**	**ΔCFI**	**LRT**	R^2	**LRT**
250	**3**	**5**	0.077	0.083	0.144	0.105	0.577	0.182	0.030
		25	0.010	0.998	0.010	0	0.399	0.026	0.032
	5	**5**	0.068	0.076	0.170	0.092	0.502	0.022	0.026
		25	0.006	0.996	0.006	0	0.406	0	0.038
1000	**3**	**5**	0.046	0.080	0.060	0.010	0.628	0	0.032
		25	0	0.996	0	0	0.438	0	0.048
	5	**5**	0.062	0.072	0.094	0.002	0.546	0	0.038
		25	0	0.996	0	0	0.366	0	0.032

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; Comb = combination of χ^2, ΔRMSEA and ΔCFI.

Table 3

Thresholds' FPR scale level - invariance scenario

			FPR scale level - thresholds						
				MG-CCFA			**MG-IRT LoR**		**MG-IRT LRT**
N	**C**	**J**	**Comb**	χ^2	**ΔRMSEA**	**ΔCFI**	**LRT**	R^2	**LRT**
250	**3**	**5**	0.078	0.078	0.225	0.211	0.660	0.189	0.036
		25	0	0.126	0	0	0.404	0.020	0.032
	5	**5**	0.088	0.088	0.224	0.202	0.527	0.020	0.036
		25	0.002	0.126	0.002	0	0.370	0	0.042
1000	**3**	**5**	0.072	0.076	0.124	0.050	0.626	0.002	0.042
		25	0	0.116	0	0	0.442	0	0.030
	5	**5**	0.102	0.102	0.152	0.072	0.528	0	0.034
		25	0	0.150	0	0	0.384	0	0.036

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; Comb = combination of χ^2, ΔRMSEA and ΔCFI.

Table 4

Loadings' FPR item level - invariance scenario

			FPR scale level- loadingss						
				MG-CCFA			**MG-IRT LoR**		**MG-IRT LRT**
N	**C**	**J**	**Comb**	χ^2	**ΔRMSEA**	**ΔCFI**	**LRT**	R^2	**LRT**
250	**3**	**5**	0.036	0.067	0.048	0.029	0.243	0.053	0.047
		25	0.001	0.312	0.001	0	0.022	0.001	0.050
	5	**5**	0.037	0.071	0.060	0.023	0.202	0.005	0.051
		25	0.001	0.326	0.001	0	0.020	0	0.049
1000	**3**	**5**	0.020	0.068	0.021	0.002	0.239	0	0.045
		25	0	0.307	0	0	0.021	0	0.057
	5	**5**	0.024	0.074	0.025	0.001	0.200	0	0.059
		25	0	0.319	0	0	0.021	0	0.047

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; Comb = combination of χ^2, ΔRMSEA and ΔCFI.

Table 5

Thresholds' FPR item level - invariance scenario

			FPR item level - thresholds						
				MG-CCFA			**MG-IRT LoR**		**MG-IRT LRT**
N	**C**	**J**	**Comb**	χ^2	**ΔRMSEA**	**ΔCFI**	**LRT**	R^2	**LRT**
250	**3**	**5**	0.052	0.070	0.059	0.067	0.236	0.053	0.051
		25	0	0.057	0	0	0.022	0.001	0.053
	5	**5**	0.059	0.076	0.076	0.075	0.194	0.010	0.048
		25	0	0.078	0	0	0.020	0	0.050
1000	**3**	**5**	0.022	0.062	0.022	0.004	0.256	0	0.048
		25	0	0.064	0	0	0.021	0	0.049
	5	**5**	0.025	0.072	0.024	0.010	0.179	0	0.040
		25	0	0.079	0	0	0.020	0	0.048

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; Comb = combination of χ^2, ΔRMSEA and ΔCFI.

Table 6

Loadings' TPR scale level - non-invariance scenario

				TPR scale level - loadings					
				MG-CCFA		MG-IRT LoR		MG-IRT LRT	
N	**C**	**J**	**%**	**small**	**large**	**small**	**large**	**small**	**large**
250	**3**	**5**	**20%**	0.079	0.063	0.177	0.154	0.048	0.043
			40%	0.107	0.163	0.183	0.242	0.054	0.079
			40%±	0.116	0.277	0.193	0.310	0.048	0.088
		25	**20%**	0.022	0.040	0.030	0.092	0.076	0.094
			40%	0.012	0.134	0.044	0.176	0.064	0.166
			40%±	0.078	0.624	0.075	0.365	0.109	0.300
	5	**5**	**20%**	0.076	0.091	0.018	0.018	0.048	0.030
			40%	0.100	0.176	0.032	0.052	0.054	0.086
			40%±	0.179	0.325	0.052	0.103	0.080	0.154
		25	**20%**	0.010	0.076	0	0.008	0.062	0.164
			40%	0.028	0.214	0.002	0.020	0.080	0.256
			40%±	0.076	0.766	0.002	0.118	0.114	0.376
1000	**3**	**5**	**20%**	0.072	0.110	0	0	0.044	0.098
			40%	0.158	0.408	0	0.032	0.084	0.322
			40%±	0.234	0.722	0.004	0.064	0.092	0.506
		25	**20%**	0.002	0.396	0	0	0.138	0.584
			40%	0.028	0.806	0	0	0.216	0.718
			40%±	0.204	1	0	0.008	0.298	0.980
	5	**5**	**20%**	0.074	0.122	0	0	0.052	0.128
			40%	0.178	0.554	0	0	0.108	0.440
			40%±	0.290	0.864	0	0.006	0.144	0.692
		25	**20%**	0.016	0.574	0	0	0.186	0.720
			40%	0.062	0.952	0	0	0.260	0.858
			40%±	0.366	1	0	0	0.398	1

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; % = percentage of items affected by DIF (± misaligned); small = small bias; large = large bias; values in red = FPR $\geq$ .10 in the invariance scenario.

Table 7

Thresholds' TPR scale level - non-invariance scenario

				TPR scale level - thresholds					
				MG-CCFA		**MG-IRT LoR**		**MG-IRT LRT**	
N	**C**	**J**	**%**	**small**	**large**	**small**	**large**	**small**	**large**
250	**3**	**5**	**20%**	0.388	0.898	0.338	0.672	0.130	0.456
			40%	0.732	0.998	0.336	0.760	0.284	0.758
			40%±	0.674	0.996	0.584	0.996	0.246	0.862
		25	**20%**	0.004	0.416	0.144	0.932	0.264	0.884
			40%	0.002	0.426	0.168	0.948	0.268	0.902
			40%±	0.012	1	0.832	1	0.468	0.996
	5	**5**	**20%**	0.512	0.984	0.078	0.458	0.104	0.504
			40%	0.836	1	0.120	0.466	0.232	0.804
			40%±	0.862	1	0.318	0.990	0.272	0.914
		25	**20%**	0.028	0.930	0.022	0.602	0.254	0.922
			40%	0.046	0.948	0.032	0.592	0.244	0.876
			40%±	0.220	1	0.612	1	0.400	0.996
1000	**3**	**5**	**20%**	0.966	1	0.026	0.478	0.550	1
			40%	1	1	0.022	0.560	0.888	1
			40%±	1	1	0.202	1	0.978	1
		25	**20%**	0.144	1	0	0.556	0.954	1
			40%	0.130	1	0	0.556	0.944	1
			40%±	0.992	1	0.626	1	1	1
	5	**5**	**20%**	0.996	1	0	0.228	0.598	1
			40%	1	1	0	0.222	0.910	1
			40%±	1	1	0.018	1	0.986	1
		25	**20%**	0.758	1	0	0.024	0.958	1
			40%	0.756	1	0	0.030	0.964	1
			40%±	1	1	0.430	1	1	1

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; % = percentage of items affected by DIF (± misaligned); small = small bias; large = large bias; values in red = FPR $\geq$ .10 in the invariance scenario.

Table 8

Loadings' TPR item level - non-invariance scenario

				TPR item level - loadings					
				MG-CCFA		**MG-IRT LoR**		**MG-IRT LRT**	
N	**C**	**J**	**%**	**small**	**large**	**small**	**large**	**small**	**large**
250	**3**	**5**	**20%**	0.018	0.033	0.004	0.004	0.061	0.064
			40%	0.047	0.082	0.055	0.088	0.067	0.116
			40%±	0.071	0.181	0.063	0.119	0.068	0.142
		25	**20%**	0.001	0.017	0.004	0.019	0.087	0.224
			40%	0.001	0.016	0.003	0.015	0.084	0.200
			40%±	0.006	0.059	0.006	0.041	0.096	0.252
	5	**5**	**20%**	0.032	0.044	0	0	0.074	0.114
			40%	0.047	0.112	0.005	0.028	0.071	0.173
			40%±	0.082	0.254	0.016	0.056	0.085	0.205
		25	**20%**	0.002	0.019	0	0.002	0.110	0.251
			40%	0.003	0.017	0	0.002	0.111	0.230
			40%±	0.009	0.086	0.001	0.012	0.111	0.303
1000	**3**	**5**	**20%**	0.016	0.046	0	0	0.098	0.178
			40%	0.060	0.177	0.001	0.013	0.109	0.338
			40%±	0.124	0.434	0.001	0.045	0.136	0.421
		25	**20%**	0	0.020	0	0	0.250	0.618
			40%	0.001	0.002	0	0	0.217	0.621
			40%±	0.002	0	0	0.001	0.261	0.707
	5	**5**	**20%**	0.024	0.074	0	0	0.112	0.206
			40%	0.084	0.251	0	0	0.156	0.454
			40%±	0.151	0.479	0	0.002	0.159	0.507
		25	**20%**	0.001	0.036	0	0	0.283	0.725
			40%	0.001	0.002	0	0	0.288	0.732
			40%±	0.007	0	0	0	0.298	0.812

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; % = percentage of items affected by DIF (± misaligned); small = small bias; large = large bias.

Table 9

Thresholds' TPR item level - non-invariance scenario

				TPR item level - thresholds					
				MG-CCFA		**MG-IRT LoR**		**MG-IRT LRT**	
N	**C**	**J**	**%**	**small**	**large**	**small**	**large**	**small**	**large**
250	**3**	**5**	**20%**	0.502	0.960	0.014	0.108	0.214	0.668
			40%	0.649	0.987	0.059	0.296	0.311	0.763
			40%±	0.568	0.983	0.237	0.545	0.294	0.780
		25	**20%**	0.001	0.195	0.010	0.372	0.349	0.886
			40%	0	0.182	0.018	0.342	0.335	0.886
			40%±	0.002	0.014	0.153	0.655	0.360	0.885
	5	**5**	**20%**	0.622	0.996	0.002	0.012	0.198	0.746
			40%	0.740	1	0.018	0.078	0.304	0.802
			40%±	0.732	0.999	0.131	0.504	0.309	0.816
		25	**20%**	0.004	0.298	0.002	0.098	0.362	0.880
			40%	0.008	0.317	0.001	0.098	0.353	0.879
			40 %±	0.042	0.053	0.100	0.526	0.339	0.875
1000	**3**	**5**	**20%**	0.962	1	0	0	0.758	1
			40%	0.985	1	0	0.056	0.869	1
			40%±	0.969	1	0.116	0.500	0.857	0.998
		25	**20%**	0.002	0	0	0.157	0.918	1
			40%	0.003	0	0	0.182	0.908	1
			40%±	0	0	0.072	0.579	0.920	1
	5	**5**	**20%**	0.992	1	0	0	0.808	1
			40%	0.998	1	0	0	0.894	1
			40%±	0.992	1	0.009	0.500	0.889	1
		25	**20%**	0.026	0.002	0	0.004	0.904	1
			40%	0.016	0.003	0	0.007	0.903	1
			40 %±	0	0.005	0.046	0.497	0.907	1

Note. MG-CCFA = Multiple-group categorical confirmatory factor analysis; MG-IRT LoR = Logistic regression with MG-IRT; MG-IRT LRT = Likelihood-ratio test with MG-IRT; N = Sample size within each group; C = Number of categories; J = Number of items; % = percentage of items affected by DIF (± misaligned); small = small bias; large = large bias.

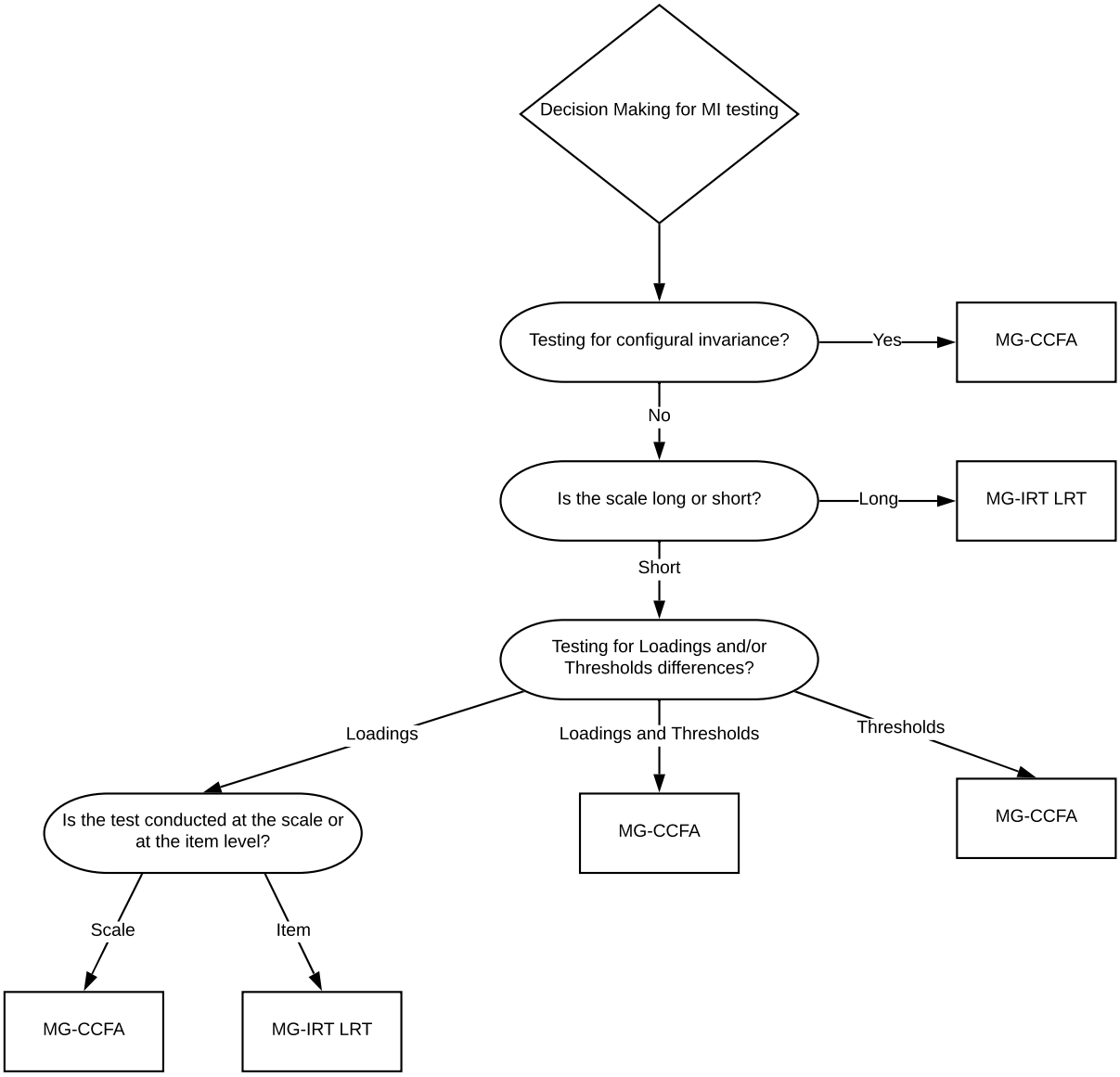

Figure 1. Flowchart, based the results of the simulation study, to provide recommendations to test meaurement invariance for ordinal data