Use of the Lagrange Multiplier Test for Assessing Measurement Invariance Under Model Misspecification

Use of the Lagrange Multiplier test for assessing measurement invariance under model misspecification

Lucia Guastadisegni

University of Bologna

Silvia Cagnone

University of Bologna

Irini Moustaki

London School of Economics and Political Science

Vassilis Vasdekis

Athens University of Economics and Business

Abstract

This paper studies the Type I error, false positive rates, and power of four versions of the Lagrange Multiplier test to detect measurement non-invariance in Item Response Theory (IRT) models for binary data under model misspecification. The tests considered are the Lagrange Multiplier test computed with the Hessian and cross-product approach, the Generalized Lagrange Multiplier test and the Generalized Jackknife Score test. The two model misspecifications are those of local dependence among items and non-normal distribution of the latent variable. The power of the tests is computed in two ways, empirically through Monte Carlo simulation methods and asymptotically, using the asymptotic distribution of each test under the alternative hypothesis. The performance of these tests is evaluated by means of a simulation study. The results highlight that, under mild model misspecification, all tests have good performance while, under strong model misspecification, the tests performance deteriorates, especially for false positive rates under local dependence and power for small sample size under misspecification of the latent variable distribution. In general, the Lagrange Multiplier test computed with the Hessian approach and the Generalized Lagrange Multiplier test have better performance in terms of false positive rates while the Lagrange Multiplier test computed with the cross-product approach has the highest power for small sample sizes. The asymptotic power turns out to be a good alternative to the classic empirical power because it is less time consuming. The Lagrange tests studied here have been also applied to a real data set.

Use of the Lagrange Multiplier test for assessing measurement invariance under model misspecification

Introduction

Item Response Theory (IRT) models are used in psychological and educational research for measuring unobserved constructs, also known as factors or latent variables, from correlated observed variables/items. The main assumptions and features of an IRT model are i) local independence among items conditional on the latent variable(s), ii) it is usually a parametric model for the probability of responding 'correctly/positively' to an item given the latent variable(s) also known as response category probability and item characteristic curve (ICC) and iii) normal distribution for the latent variable(s) (Bartholomew, Knott, and Moustaki 2011). As with any statistical model, some of the above assumptions may be violated. The Likelihood-Ratio, the Wald, and the Lagrange Multiplier or score (LM) test statistics (Cox and Hinkley 1979) check model fit and they are asymptotically equivalent. Differently from the Likelihood-Ratio and the Wald test, the LM test only requires the computation of the restricted estimator (model under the null hypothesis). The LM test can be very convenient in IRT models, where multiple model violations (e.g. local dependence, non-normality of latent distribution, etc.) can occur (Fox and Glas 2005). The LM test does not need the estimation of an alternative model for each one of these violations. Moreover, there is model violation, such as differential item functioning (DIF), that requires testing items sequentially (Glas 1998). The LM test does not require new parameter estimates for every tested item, making it computationally less intensive, especially in long tests. For these reasons, the LM test is used in IRT to detect DIF (Glas 1998, Fox and Glas 2005), local dependence (LD) (Glas 1999, Glas and Falcón 2003, Fox and Glas 2005, Kim, De Ayala, Ferdous, and Nering 2011, Liu and Thissen 2012, Liu and Maydeu-Olivares 2013, Liu and Thissen 2014, van der Linden and Glas 2010, Oberski, van Kollenburg, and Vermunt 2013) and deviation from the parametric model (i.e. ICC) (Glas 1999, Glas and Falcón 2003, Ranger and Kuhn 2012).

The LM test depends on the Fisher information matrix. Different approximations

of this matrix lead to different test performances. Accurate results for the LM test can be obtained by considering the expected Hessian and cross-product matrix, as shown in Liu and Maydeu-Olivares (2013), but they are unfeasible in long tests. For this reason, the observed versions of these matrices are preferred for the computation of the LM test. Some authors (Glas 1998, Oberski et al. 2013) use the observed Hessian matrix, that we denote with LM(H), and others (Liu and Maydeu-Olivares 2013, Liu and Thissen 2012, Liu and Thissen 2014) the observed cross-product matrix, that we denote with LM(CP). Falk and Monroe (2018) compare both approaches. The LM(CP) test shows more inflated Type I error rates than the LM(H) test, especially with long tests and small sample size, but it is fast to compute (Liu and Thissen 2012, Liu and Maydeu-Olivares 2013, Liu and Thissen 2014, Falk and Monroe 2018). In some works, the LM test statistic is applied in the case of model misspecification under the null and the alternative hypotheses, showing a good performance when the amount of model misspecification is overall small (Glas and Falcón 2003, Falk and Monroe 2018, Guastadisegni, Cagnone, Moustaki, and Vasdekis forthcoming). Different versions of the LM test are also derived under model misspecification (White 1982, Boos 1992). White (1982) proposes the Generalized Lagrange Multiplier (LM(S)) test, whose expression involves the sandwich variance and covariance matrix. Similarly Boos (1992) derives a Generalized Score (GS) test for least squares, robust M-estimation, and quasi-likelihood estimation methods that is equivalent to the LM(S) test when maximum likelihood (ML)-based methods are used. The Generalized Jackknife Score (GS(J)) test is a version of the GS test, derived under model misspecification, where the covariance matrix of the score is computed using the Jackknife estimates (J. Rao, Scott, and Skinner 1998). The GS(J) test has not been studied in the IRT context. As far as we know, the LM(S) test is studied only by Falk and Monroe (2018) and Guastadisegni et al. (forthcoming). Falk and Monroe (2018) compare the performance of the LM(S), LM(CP), and LM(H) tests for a single omitted cross-loading and Guastadisegni et al. (forthcoming) compute the empirical and asymptotic power of the LM(S) and LM(H) tests to assess measurement invariance under misspecification of the latent variable

distribution, without studying the Type I error/false positive rates of these two tests. Differently from these works, we assess measurement invariance considering a more general framework, where the model misspecification is due to local dependence among items and different non-normal latent variable distributions.

In the case of a one factor model, an item is measurement invariant if the conditional distribution of the item given the latent variable is independent of group membership identified by an external group variable (e.g. sex, age, country) (Mellenbergh 1982,1983). An item is measurement non-invariant (also known as DIF), if it measures different abilities for different group memberships. In this case, the expected score of the item differs in the subgroups for the same level of the latent variable. Measurement invariance can be studied either in a multiple-group analysis setup (Jöreskog 1971) or with the Multiple Indicator Multiple Causes (MIMIC) model (Jöreskog and Goldberger 1975). The model allows direct and indirect effects of a binary group covariate on the probability of giving a 'correct/positive' response to an item and on the latent variable respectively.

The contribution of this paper is twofold. First, we assess item measurement invariance under model misspecification, using four versions of the LM test. The four versions differ in the form of the covariance matrix of the estimators. Mainly, the Hessian estimator (LM(H)), the cross-product estimator (LM(CP)), the sandwich estimator (LM(S)), and the Jackknife estimator (GS(J)) are discussed and studied here. Second, we compute the power of the LM(H), LM(CP), and LM(S) tests in two ways, empirically through Monte Carlo simulation methods and asymptotically using the distribution of each test under the alternative hypothesis, which depends on a non-centrality parameter often difficult to compute (Gudicha, Schmittmann, and Vermunt 2017). The non-centrality parameter is approximated using the procedure derived by Gudicha et al. (2017) for the Wald and Likelihood-Ratio tests and it is applied in Guastadisegni et al. (forthcoming) to the LM(H) and LM(S) tests under misspecification of the latent variable distribution. We extend this method to the case of local dependence and to the LM(CP) test.

Through an extensive simulation study, we compare the performance of the different versions of the LM tests in terms of Type I error rate, false positive rate, and empirical and asymptotic power, varying the type and the misspecification level and considering single and multiple parameter hypotheses tests for measurement invariance. Moreover, we illustrate the use of these tests to a real data set.

The paper is organized as follows. First, we present the MIMIC model with covariate effects. Second, we describe the four versions of the LM tests and the procedure to estimate the asymptotic power for the LM(H), LM(CP), and LM(S) tests. Next, we present a Monte Carlo simulation study and the results from the real data analysis. Finally, some concluding remarks are presented and discussed.

The MIMIC model for binary data

Let us denote by $y_1, ..., y_p$ a set of observed binary variables/items, by z the latent variable, and by x a binary variable such as sex, country, or any other group variable. Given n individuals, the i-th subject belongs to either the focal or the reference group when $x_i = 1$ or $x_i = 0$ respectively. To test for item(s)' measurement invariance, we consider the MIMIC model with the group variable x affecting both the item(s) y and the latent variable z. Group differences can be present only on the item intercept (uniform-DIF) or simultaneously on the item intercept and slope (non-uniform DIF) (Glas 1998, Fox and Glas 2005). The response probability for the i-th individual to the j-th item is modelled using a logistic model (measurement model) where the model for the latent variable is a linear model (structural model) defined by:

$$\begin{aligned} P(y_{ij} = 1|z_i, x_i) = \pi_{ij}(z_i, x_i) = \frac{\exp\left(\alpha_{0j} + \alpha_{1j} z_i + \gamma_{1j} x_i + \gamma_{2j} x_i z_i\right)}{1 + \exp\left(\alpha_{0j} + \alpha_{1j} z_i + \gamma_{1j} x_i + \gamma_{2j} x_i z_i\right)} \\ z_i = \beta x_i + \epsilon_i \qquad \epsilon \sim N(0, 1) \end{aligned} \tag{1}$$

where $i = 1, ..., n$ and $j = 1, ..., p$. Under non-uniform DIF, the intercept and factor loading parameters are $(\alpha_{0j}, \alpha_{1j})$, and $(\alpha_{0j} + \gamma_{1j}, \alpha_{1j} + \gamma_{2j})$ for the reference and focal groups respectively (Glas 1998). The parameter β allows the mean of the latent variable z to be different in the two groups, although it is set to $N(0, 1)$ in the reference group

for identification purposes. For a random sample of size n the log-likelihood is:

$$l(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\mathbf{y}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \int \prod_{j=1}^{p} \pi_{ij}(z_i, x_i)^{y_{ij}} (1 - \pi_{ij}(z_i, x_i))^{1-y_{ij}} \phi(z_i \mid x_i) dz_i, \quad (2)$$

where $\boldsymbol{\theta}$ is the vector of the unknown parameters and the model assumes conditional/local independence among the items. Equation (2) is maximized using either an expectation–maximization (EM) algorithm (Bock and Aitkin 1981) or a direct maximization, such as the Newton-Raphson algorithm (Skrondal and Rabe-Hesketh 2004).

Uniform and non-uniform DIF for an item y_j is assessed by testing the statistical significance of the parameters γ_{1j} and (γ_{1j},γ_{2j}) respectively. We consider situations where the parameters γ_{1j} or $(\gamma_{1j}, \gamma_{2j})$ are fixed to zero and to constants different from zero under the null hypothesis. Moreover, the performance of the LM tests is assessed under violations of local independence and normality distribution of the latent variable.

Lagrange Multiplier tests

The classical Lagrange Multiplier test

The LM test (C. R. Rao 1948) evaluates the statistical significance of imposed restrictions on model parameters. We consider a sample $\mathbf{y}_1, ..., \mathbf{y}_n$ from a model $f(\mathbf{y}, \boldsymbol{\theta})$. The true parameter vector is denoted by $\boldsymbol{\theta}_0$. Let $\boldsymbol{\theta}_0$ be divided into two sub-vectors $\boldsymbol{\theta}_0' = (\boldsymbol{\theta}_{01}', \boldsymbol{\theta}_{02}')$. $\boldsymbol{\theta}_{01}$ includes the intercept parameters $(\alpha_{0j}, j = 1 \ldots, p)$ and factor regression coefficients $(\alpha_{1j}, j = 1 \ldots, p)$. When uniform-DIF is assessed, $\boldsymbol{\theta}_{02}$ includes the parameters γ_{1j} and when non-uniform DIF is assessed, $\boldsymbol{\theta}_{02}$ includes γ_{1j} and γ_{2j}, where $j = 1 \ldots, p$. The hypotheses H_0 and H_1 can be formalized as follows:

$$H_0 : \boldsymbol{\theta}_{02}' = \mathbf{c} \quad vs \quad H_1 : \boldsymbol{\theta}_{02}' \neq \mathbf{c}, \quad (3)$$

where $\mathbf{c}$ is a vector of constants.

The LM statistic is (C. R. Rao 1948):

$$LM = S(\tilde{\boldsymbol{\theta}})' A_n(\tilde{\boldsymbol{\theta}})^{-1} S(\tilde{\boldsymbol{\theta}}), \tag{4}$$

where $\tilde{\boldsymbol{\theta}}' = (\tilde{\boldsymbol{\theta}}_1', \mathbf{c})$ denotes the restricted maximum likelihood estimates of the parameters $\boldsymbol{\theta}$, $S(\tilde{\boldsymbol{\theta}}) = \frac{\partial \ln l(\boldsymbol{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the vector of score functions evaluated at $\tilde{\boldsymbol{\theta}}$, and $A_n(\tilde{\boldsymbol{\theta}}) = -E\left[\frac{\partial^2 l(\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$ is the Fisher information matrix evaluated at $\tilde{\boldsymbol{\theta}}$. Given that the part of the score vector evaluated in $\tilde{\boldsymbol{\theta}}_{01}$ is $\mathbf{0}$, the LM statistic given in (4) is reduced to:

$$LM = S_2(\tilde{\boldsymbol{\theta}}) A_n^{22}(\tilde{\boldsymbol{\theta}})^{-1} S_2(\tilde{\boldsymbol{\theta}}), \tag{5}$$

where $S_2(\tilde{\boldsymbol{\theta}})$ is a subset of $S(\tilde{\boldsymbol{\theta}})$ that corresponds to the parameters $\boldsymbol{\theta}_{02}$ evaluated at $\tilde{\boldsymbol{\theta}}$ and $A_n^{22}(\tilde{\boldsymbol{\theta}})$ is a block of the partitioned Fisher information matrix computed as (Engle 1984)

$$A_n^{22} = A_{n22} - A_{n21} A_{n11}^{-1} A_{n12}, \tag{6}$$

and evaluated at $\tilde{\boldsymbol{\theta}}$. The partition of A_n into $A_{n22}, A_{n21}, A_{n11}, A_{n12}$ is derived from the partition of $\boldsymbol{\theta}_0'$ into $(\boldsymbol{\theta}_{01}', \boldsymbol{\theta}_{02}')$.

Two different versions of the LM test are studied here depending on which matrix is used for estimating $A_n(\tilde{\boldsymbol{\theta}})$. The Hessian approach (LM(H)), uses the observed Hessian matrix given by

$$\hat{A}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2 l_i(\mathbf{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \tag{7}$$

whereas the cross-product approach (LM(CP)), uses the observed cross-product matrix

$$\hat{B}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \ln l_i(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln l_i(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{8}$$

Under correct model specification, $\hat{A}_n(\boldsymbol{\theta}) = \hat{B}_n(\boldsymbol{\theta})$ (White 1982) and the LM(H) and LM(CP) tests are equivalent.

Under a correctly specified likelihood and under H_0, the LM test statistic,

computed with the Hessian and cross-product approaches, is asymptotically distributed as a χ^2_r, with degrees of freedom (r) equal to the dimension of $\boldsymbol{\theta}_{02}$.

To compute the local asymptotic power of the LM test, a standard approach is to consider a set of local alternatives close to the null value for large n, $H_1 : \boldsymbol{\theta}_{02} = \boldsymbol{c} + \frac{\boldsymbol{\xi}}{\sqrt{n}}$, where $\boldsymbol{\xi}$ is an arbitrary vector with the same dimension of $\boldsymbol{\theta}_{02}$ (Boos and Stefanski 2013). When the model defined under H_1 is true, the LM test is asymptotically distributed as a non-central chi-square that depends on two parameters, namely the degrees of freedom (equal to the dimension of $\boldsymbol{\theta}_{02}$), and a non-centrality parameter λ given by (Cox and Hinkley 1979):

$$\lambda = \frac{1}{n}\boldsymbol{\xi}' A_n^{22}(\boldsymbol{\theta_0})\boldsymbol{\xi} \tag{9}$$

The asymptotic power is computed as $P(\chi^2_r(\lambda) > \chi^2_r(1-\alpha))$.

Approximation procedure for the asymptotic power. The asymptotic distribution of the LM test as a non-central chi-square with non-centrality parameter in equation (9) holds when the model defined under the set of local alternatives is true, i.e. when the model under the null hypothesis is barely incorrect for large n (see Agresti 2002, Reiser 2008). In practice, it is often reasonable to adopt an alternative hypothesis for fixed and finite n (Agresti 2002), as $H_1 : \boldsymbol{\theta}_{02} = \boldsymbol{c} + \boldsymbol{\xi}$, or to use hypotheses as in (3) (Gudicha et al. 2017). Here, we consider the approximation procedure for the asymptotic power derived by Gudicha et al. (2017) for the Likelihood-Ratio and the Wald tests. This procedure is extended to the LM(H) test in Guastadisegni et al. (forthcoming). The method can also be used for the LM(CP) test and can be summarized in the following steps:

1. From the model defined under the alternative hypothesis, create a large data set (e.g. $N = 10000$ observations).

2. Fit the model under H_0 to the data generated under step 1.

3. Take the value of the LM(H)/LM(CP) statistic as the estimate of the non-centrality parameter λ (Satorra 1989, Bollen 1989).

4. Compute the non-centrality parameter for a sample of size 1 equal to $\lambda_1 = \frac{\lambda}{N}$.

5. The non-centrality parameter for a sample of size n is $\lambda_n = n\lambda_1$.

The asymptotic power of the LM(H)/LM(CP) test can be determined by comparing the λ_n obtained in step 5 with the tabled values of the non-central chi-square with *df* corresponding to the number of parameters constrained under H_0 and significance level α (Bollen 1989).

The Generalized Lagrange Multiplier test

Consider a sample $\mathbf{y}_1, ..., \mathbf{y}_n$ from a model with true density $g(\mathbf{y})$, that assumes either local dependence among the items or a non-normal distribution of the latent variable. The model with density $f(\mathbf{y}; \boldsymbol{\theta})$, which assumes both local independence among the items and a normal distribution of the latent variable, is erroneously assumed to be the true model for the data and it is used for ML analysis. If the assumptions A1-A6 (pp: 2-6, White 1982), that ensure the existence, consistency, asymptotic normality, and identifiability of the Quasi-ML estimator, are fulfilled, the parameter vector $\hat{\boldsymbol{\theta}}_n$, which maximizes the log-likelihood function based on model $f(\mathbf{y}; \boldsymbol{\theta})$, converges in probability to $\boldsymbol{\theta}_*$, the parameter vector that minimizes the Kullback-Leibler information criterion. Moreover, the covariance matrix of $\hat{\boldsymbol{\theta}}_n$, based on n observations, is the so-called sandwich estimator given by $\hat{C}_n(\hat{\boldsymbol{\theta}}_n) = \hat{A}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\hat{B}_n(\hat{\boldsymbol{\theta}}_n)\hat{A}_n^{-1}(\hat{\boldsymbol{\theta}}_n)$, where the matrix $\hat{A}_n$ and $\hat{B}_n$ are the observed Hessian matrix and the observed cross-product matrix defined in formulas (7) and (8) respectively and evaluated at $\hat{\boldsymbol{\theta}}_n$.

Under model misspecification, the null and the alternative hypotheses are now specified in terms of $\boldsymbol{\theta}_*$. Let $\boldsymbol{\theta}_*$ be divided in two sub-vectors $\boldsymbol{\theta}'_* = (\boldsymbol{\theta}'_{*1}, \boldsymbol{\theta}'_{*2})$. To test for uniform and non-uniform DIF, the parameters $\boldsymbol{\theta}'_{*1}, \boldsymbol{\theta}'_{*2}$ are grouped as in The classical Lagrange Multiplier test section. The hypotheses in (3) can be formalized as follows:

$$H_0 : \boldsymbol{\theta}'_{*2} = \mathbf{c} \quad vs \quad H_1 : \boldsymbol{\theta}'_{*2} \neq \mathbf{c}, \tag{10}$$

where **c** is a vector of constants.

The Generalized Lagrange Multiplier test is defined as (White 1982, Engle 1984):

$$LM(S) = S_2(\tilde{\boldsymbol{\theta}}_n)' \hat{A}_n^{22}(\tilde{\boldsymbol{\theta}}_n)^{-1} \hat{C}_{n22}(\tilde{\boldsymbol{\theta}}_n)^{-1} \hat{A}_n^{22}(\tilde{\boldsymbol{\theta}}_n)^{-1} S_2(\tilde{\boldsymbol{\theta}}_n), \tag{11}$$

where $\hat{A}^{22}(\tilde{\boldsymbol{\theta}}_n)$ is computed as in (6) replacing A_n with $\hat{A}_n$, evaluated at $\tilde{\boldsymbol{\theta}}_n$ and $\hat{C}_{n22}(\tilde{\boldsymbol{\theta}}_n)$ is the part of the matrix $\hat{C}_n$ corresponding to $\boldsymbol{\theta}'_{*2}$, evaluated at $\tilde{\boldsymbol{\theta}}_n$. Under H_0, LM(S) is distributed as a χ^2_r, with degrees of freedom r equal to the dimension of $\boldsymbol{\theta}_{*2}$. If the model is correctly specified, the statistic LM(S) is equal to the LM test, computed both with the Hessian or the cross-product approach (White 1982).

As before, the local asymptotic power of the LM(S) test is obtained by considering a set of local alternatives given by $H_1 : \boldsymbol{\theta}_{*2} = \boldsymbol{c} + \frac{\boldsymbol{\xi}}{\sqrt{n}}$, where $\boldsymbol{\xi}$ is an arbitrary vector of dimension $\boldsymbol{\theta}_{*2}$. Under H_1, LM(S) converges in distribution to a $\chi^2_r(\lambda)$, with degrees of freedom r equal to the dimension of $\boldsymbol{\theta}_{*2}$ and λ is the non-centrality parameter given by (Bera et al. 2020):

$$\lambda = \frac{1}{n} \xi' A_n^{22'} (B_{n22} - A_{n21} A_{n11}^{-1} B_{n12} - B_{n21} A_{n11}^{-1} A_{n12} + A_{n21} A_{n11}^{-1} B_{n11} A_{n11}^{-1} A_{n12})^{-1} A_n^{22} \xi \tag{12}$$

where $A_{n11}, A_{n12}, A_{n21}$ are the blocks of the expected Fisher information matrix A_n and $B_{n11}, B_{n12}, B_{n21}, B_{n22}$ of the expected cross-product matrix B_n, derived from the partition of $\boldsymbol{\theta}'_*$ into $(\boldsymbol{\theta}'_{*1}, \boldsymbol{\theta}'_{*2})$. A_n^{22} is computed as in (6). All matrices in formula (12) are evaluated at $\boldsymbol{\theta}^*$. The asymptotic power estimation method described in the Approximation procedure for the asymptotic power section is used here to estimate the asymptotic power for the LM(S) test. In step 3, the LM(S) statistic is taken as the estimate of the non-centrality parameter (the proof of this result can be found in Satorra 1989). Moreover, the model fitted under H_0 at step 2 is assumed to be misspecified. Under correct model specification the LM(S) and the LM(H)/LM(CP) tests have the same non-centrality parameter and, consequently, the same asymptotic power.

The Jackknife Generalized Score test

When ML-based methods are used, the LM(S) test derived by White (1982) is equivalent to the GS test derived by Boos (1992) under model misspecification and valid under different types of estimation methods, such as least squares, quasi-ML, and robust M-estimation. The Generalized Score test for the hypothesis testing given in (10) is:

$$GS = S_2(\tilde{\boldsymbol{\theta}})' V_{S_2}^{-1}(\tilde{\boldsymbol{\theta}}) S_2(\tilde{\boldsymbol{\theta}}), \tag{13}$$

where $S_2(\tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}$ are defined similarly as in The Generalized Lagrange Multiplier test section, but S_2 does not necessarily come from the derivative of a log-likelihood because it depends on the estimation method chosen. $V_{s_2}(\tilde{\boldsymbol{\theta}})$ is the covariance matrix of S_2, evaluated at $\tilde{\boldsymbol{\theta}}$.

When likelihood-based methods are used, $V_{s_2}(\tilde{\boldsymbol{\theta}})$ is equal to $\hat{A}_n^{22}(\tilde{\boldsymbol{\theta}})\hat{C}_{n22}(\tilde{\boldsymbol{\theta}})\hat{A}_n^{22}(\tilde{\boldsymbol{\theta}})$ and formulas (13) and (11) are equivalent. Under H_0, the GS test is distributed as a χ_r^2, where r are the *df* equal to the dimension of $\boldsymbol{\theta}_{*2}$.

J. Rao et al. (1998) proposed a version of the Generalized Score test in a general estimating equations framework (Godambe and Thompson 1986) for a stratified multistage sampling design, based on a consistent Jackknife estimator of $V_{S_2}(\tilde{\boldsymbol{\theta}})$. We use the test proposed by J. Rao et al. (1998), for independent and identically distributed (i.i.d.) observations and maximum likelihood estimation methods and we refer to this test as the Jackknife Generalized Score (GS(J)) test. The GS(J) test is given in formula (13), where $V_{S_2}(\tilde{\boldsymbol{\theta}})$ is estimated with the delete-1 Jackknife method as:

$$\hat{V}_{s_2}(\tilde{\boldsymbol{\theta}}_n) = \frac{n}{n-1}\sum_{i=1}^{n}(\tilde{S}_{2(i)} - \tilde{S}_2)(\tilde{S}_{2(i)} - \tilde{S}_2)'. \tag{14}$$

$\tilde{S}_{2(i)}$ is the score function computed by removing the i-th observation and evaluated at $\tilde{\boldsymbol{\theta}}_{n(i)}$, (i.e. the ML estimate obtained by maximizing the score function without the i-th observation), and $\tilde{S}_2$ is the score function of the original sample evaluated at $\tilde{\boldsymbol{\theta}}_n$. Shao (1992) proved the consistency of the Jackknife method for a parameter estimator θ for i.i.d. responses, while J. Rao et al. (1998) gave a sketch of the proof of the consistency

of the Jackknife score variance estimator for basic survey weights.

Simulation study

We study the performance of the LM(H), LM(CP), LM(S), and GS(J) test statistics under no misspecification and misspecification either due to local dependence or in the latent variable distribution. Since the main focus of this work is the case of model misspecification, the results under correct model specification are reported in the Supplementary material. Under a correct model specification, data are generated from the two-Parameter Logistic (2-PL) model (Birnbaum 1968) with a linear structural model. When the model is correctly specified, we find results in line with the literature. In particular, the LM(CP) test shows inflated Type I error rates whereas the LM(H) and LM(S) tests have simulated Type I error rates quite close to the nominal level α and similar power. Moreover, the power of the tests increases with the sample size and the number of items. Similar results are found by Liu and Maydeu-Olivares (2013), Liu and Thissen (2014), and Falk and Monroe (2018).

In the Violation of local independence and the Misspecification of the latent variable distribution sections, uniform and non-uniform DIF are studied in the simulation as well as single and multiple parameter hypotheses. The performance of the GS(J) test is evaluated in a separate simulation study in The study on the GS(J) test section.

We consider the following simulation conditions: number of items ($p = 10, 20$) $\times$ sample size ($n = 200, 500, 1000$)$\times$ test statistic ($LM(H), LM(CP), LM(S)$). To evaluate the asymptotic behaviour of the tests, in some of the cases, $n = 5000$ is considered. In some cases, the asymptotic power is computed in addition to the empirical power. Direct maximization through the Newton-Raphson method is used to obtain the ML-estimates under the null hypothesis and numerical derivatives are used to compute the Hessian and cross-product matrices.

The optimization is conducted in R with the function “optim”, and numerical derivatives are obtained with the “NumDeriv” R package. In all the simulation

scenarios, $N = 500$ replications are considered and the nominal level α is fixed to 0.05. Only for the results under correct model specification, and reported in the Supplementary Material, do we consider $N = 200$.

Under model misspecification, in hypothesis testing we should account for the true data generating value θ_0 and for the parameter value θ_* as follows:

- when $H_0 : \theta_* = c$, provided that $\theta_0 = c$ and $\theta_* = c$, the Type I error rate is obtained. The null hypothesis is true under model misspecification and the parameter is correctly fixed to its data generating value.

- when $H_0 : \theta_* = c$, provided that $\theta_0 = c$ and $\theta_* \neq c$, the false positive rate is obtained. The null hypothesis is not true under model misspecification, but the parameter is correctly fixed to its data generating value. Some authors, such as Green, Thompson, and Babyak (1998), consider the rejections of parameter fixed to its data generating value as Type I error instead of false positive rate, even under model misspecification. For this reason, we expect the tests to have false positive rates close to the nominal level α if they have good performance.

- when $H_0 : \theta_* = c$, provided that $\theta_0 \neq c$ and $\theta_* \neq c$, the power is obtained. The null hypothesis is not true under model misspecification and the parameter is not fixed to its data generating value.

- the case $H_0 : \theta_* \neq c$, provided that $\theta_0 \neq c$ and $\theta_* = c$, is not examined in this study.

To estimate the unknown parameters $\boldsymbol{\theta}_*$, we fit the unconstrained model under hypothesis H_1 to a sample of 5000 observations generated from the true model. Under model misspecification we always study the false positive rates instead of the Type I error rates ($\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_*$). Non-valid statistics, for example negative statistics, are excluded from the analysis. The Type I error, false positive, and power rates are computed as $\hat{p} = \sum_{l=1}^{N_v} \frac{I(T_l \geq c)}{N_v}$, where N_v is the number of valid statistics out of the number of replications, I is an indicator function, T_l is the value of the test statistic evaluated in the l-th replication and c is the theoretical asymptotic critical value corresponding to

the 95th percentile of the χ^2_{df} distribution, with degrees of freedom equal to the number of constrained parameter(s) under H_0. The confidence interval (CI) of each rate $\hat{p}$ is computed as $\hat{p} \pm 1.96\sqrt{\frac{0.05(1-0.05)}{N_v}}$.

Violation of local independence

Conditional dependence among certain items is introduced in the data generating model via a common individual specific random variables u in the logistic measurement model. Data are generated from the following model:

$$\begin{aligned} &logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i, \qquad i = 1, ..., n \qquad j = 1, ..., d,\ 1 \leq d \leq p \\ &logit(\pi_{iJ}) = \alpha_{0J} + \alpha_{1J} z_i + u_i, \qquad J = d+1, ..., p \qquad u \sim N(0, \sigma_u^2) \\ &z_i = \beta x_i + \epsilon_i \qquad \epsilon \sim N(0,1) \end{aligned} \tag{15}$$

Both for $p = 10$ and for $p = 20$, the intercept parameters are generated from a multivariate log-normal distribution with mean 0 and standard deviation (SD) 0.1, the slope parameters are generated from a multivariate log-normal distribution with mean 0 and SD 0.5, the values of the covariate x are generated from a Bernoulli distribution with success probability equal to 0.7, and the residuals ϵ are generated from a standard normal distribution. The parameter β is fixed to 0.9. The random effects u induce the local dependence among the items $y_{d+1}, ..., y_p$. The percentages of local dependent items considered in the simulations are 20% and 50%. For example, when $LD = 20\%$ and $p = 10$, two items are local dependent. Also, σ_u^2 influences the amount of misspecification in the simulation study. The random effects are generated from a normal distribution with mean 0 and three different values of σ_u^2, 0.25, 1, and 2.25. In the data generating model there is absence of uniform and non-uniform DIF.

To test for non-uniform DIF under model misspecification, we consider the

following unconstrained model:

$$
\begin{aligned}
&logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i, \qquad i = 1, ..., n \qquad j = 1, 2, ..., K \qquad 1 \leq k \leq p \\
&logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i + \gamma_{1j} x_i + \gamma_{2j} x_i z_i, \qquad j = k+1, ..., p \\
&z_i = \beta x_i + \epsilon_i, \qquad \epsilon \sim N(0, 1),
\end{aligned}
\tag{16}
$$

where items $(k+1, ..., p)$ are tested for measurement invariance. In the case of uniform DIF, equation (16) does not include the parameter γ_{2j} on the items $k+1, ..., p$.

In our simulations, the model fitted to the data is given in (16) with parameters γ_{1j} and γ_{2j} fixed to constant values. The false positive rates are studied using hypotheses A, B, and C and the empirical power using hypotheses D, E, and F. The asymptotic power is studied for scenario D.

A $H_0 : \gamma_{1j*} = 0 \qquad vs \qquad H_1 : \gamma_{1j*} \neq 0,$

This implies that one item is tested for uniform DIF.

B $H_0 : \boldsymbol{\gamma}'_{1*} = \mathbf{0} \qquad vs \qquad H_1 : \boldsymbol{\gamma}'_{1*} \neq \mathbf{0},$

where $\boldsymbol{\gamma}'_{1*}$, is a 5×1 vector (i.e. five items are tested for uniform DIF).

C $H_0 : (\gamma_{1j*}, \gamma_{2j*}) = \mathbf{0} \qquad vs \qquad H_1 : (\gamma_{1j*}, \gamma_{2j*}) \neq \mathbf{0},$

One item is tested for non-uniform DIF.

D $H_0 : \gamma_{1j*} = 0.7 \qquad vs \qquad H_1 : \gamma_{1j*} \neq 0.7,$

One item is tested for uniform DIF.

E $H_0 : \boldsymbol{\gamma}'_{1*} = \mathbf{c} \qquad vs \qquad H_1 : \boldsymbol{\gamma}'_{1*} \neq \mathbf{c}$, where $\boldsymbol{c} = (0.7, 0.7, 0.7, 0.7, 0.7),$

Five items are tested for uniform DIF.

F $H_0 : (\gamma_{1j*}, \gamma_{2j*}) = \boldsymbol{c} \qquad vs \qquad H_1 : (\gamma_{1j*}, \gamma_{2j*}) \neq \boldsymbol{c}$, where $\mathbf{c} = (0.7, 1),$

One item is tested for non-uniform DIF.

Table 1 presents the false positive rates for the LM(H), LM(CP), and LM(S) tests under local dependence for scenarios A, B, and C.

In the majority of cases, we can see that when the variance of the random effect is low ($\sigma_u^2 = 0.25$), the false positive rates of the LM(H) and LM(S) tests are quite close to the nominal level $\alpha = 5\%$, while the LM(CP) test rejects more often than expected. With the increase of model misspecification ($\sigma_u^2 = 1$ and $LD = 50\%$, $\sigma_u^2 = 2.25$ and $LD = 20\%, 50\%$) the false positive rates increase with the sample size and there are no significant differences in tests behaviour between 10 and 20 items. It is evident that the false positive rates are dramatically affected by the variance of the random effect and the number of items that are conditionally dependent. Moreover, the LM(CP) test has the most inflated false positive rates under all conditions of the study, while no improvement has been found when using the LM(S) test. Both LM(S) and LM(H) show a very similar behaviour under all scenarios.

Table 2 presents the empirical and asymptotic power for the LM(H), LM(CP), and LM(S) tests under local dependence for scenario D.

Overall, there are some numerical differences between the asymptotic and empirical power that decrease with the increase in the number of items and the sample size. It is worth noting that the behaviour of the empirical and asymptotic power is the same. Indeed, according to both methods, LM(CP) has the highest power and LM(H) and LM(S) have a very similar power under all conditions. The empirical and asymptotic power increases with both the sample size and the number of items. Since there are no substantial differences between the two procedures, only the empirical power is computed for scenarios E and F. Table 3 presents the empirical power for the LM(H), LM(CP), and LM(S) tests under local dependence for scenarios E and F.

Under the multiple parameters scenarios (E and F) and small sample sizes ($n = 200$), the LM(S) test has the lowest power. Moreover, under all scenarios and for small sample size, LM(H) and LM(CP) have similar power whereas, in the majority of cases for large sample sizes, all tests reach the same power. Thus, the power seems less affected by the degree of local dependence compared to the the false positive rate and it increases with both the sample size and the number of items. Moreover, in terms of power, LM(CP) has the best performance because it has the highest power under most

simulation conditions and it produces valid results for all replications. It is worth noting that, under scenarios E and F, in some cases the LM(H) test produces non-valid results, ranging from 0.2% to 22.4% of the replications, where the highest percentages correspond to small sample sizes, $\sigma_u^2 = 2.25$ and $LD = 50\%$.

Misspecification of the latent variable distribution

The data are generated from the following model:

$$\begin{aligned} &logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i \\ &z_i = \beta x_i + \epsilon_i, \qquad i = 1, ..., n \qquad j = 1, 2, ..., p \end{aligned} \tag{17}$$

Three different distributions are assumed for the latent variable. Namely, the error term is generated from a mixture of normals as $\epsilon \sim f(\epsilon) = 0.3N(-1.5, 0.2) + 0.7N(1, 0.4)$ and also from a skew-normal distribution with parameter $\kappa = 1, 3$. The probability density function of a skew-normal with skewness parameter κ is the following (Azzalini 1985):

$$\phi(\epsilon; \kappa) = 2\phi(\epsilon)\Phi(\epsilon; \kappa)$$

where ϕ and Φ are the standard normal density and distribution function, respectively. The parameter κ can take values from $-\infty$ to $+\infty$ and for $\kappa = 0$ reduces to a standard normal distribution.

Intercepts (α_{0j}), factor coefficients (α_{1j}), regression coefficient (β), and group variable x are generated as in the Violation of local independence section. Similarly here, we consider the model in equation (16) as the unconstrained model. The simulation scenarios of the Violation of local independence section are considered here to study the false positive rates and the empirical power of the tests. As before, the asymptotic power is studied for scenario D.

Table 4 reports the false positive rates for the LM(H), LM(CP), and LM(S) tests under misspecification of the latent variable distribution for scenarios A, B, and C.

The misspecification of the latent variable distribution in the case of a mixture of

normals does not affect the false positive rates of the LM(H) and LM(S) tests, whereas the LM(CP) test has inflated false positive rates, especially under scenarios B and C. When $\epsilon \sim SN(1)$, only the LM(S) test never shows inflated false positive rates, even if it rejects less than it should for small sample sizes and 10 items. The performance of the tests deteriorates with the increase of skewness from $\kappa = 1$ to $\kappa = 3$. For some of our simulation scenarios, the LM(H) and the LM(CP) tests have inflated false positive rates and the LM(S) test rejects less than expected. When ϵ is distributed as a skew-normal under all scenarios, the LM(H) test produces a considerable number of non-valid results, ranging from 0.2% to 43.4% of the replications. The number of non-valid LM(H) statistics increases with the skewness of the latent variable distribution and for small sample sizes.

Table 5 presents the empirical and asymptotic power for LM(H), LM(CP), and LM(S) tests under misspecification of the latent variable distribution for scenario D.

Overall, the numerical differences between the asymptotic and empirical power are small. As in the case of local dependence, the empirical and asymptotic power give the same information. For scenario D and large sample sizes, the power of all tests is not affected by the latent variable having a mixture of normal distributions. When $\epsilon \sim SN(1)$, LM(CP) has the highest power while LM(H) and LM(S) have a very similar power. When $\epsilon \sim SN(3)$, the power is lower for all tests, especially for LM(S) and small sample sizes, and LM(H) produces a considerable number of non-valid results for small sample size (11.6% of the replications). Since there are no substantial differences between the two procedures, only the empirical power is computed for scenarios E and F.

Table 6 presents the power for LM(H), LM(CP), and LM(S) tests under misspecification of the latent variable distribution for scenarios E and F.

Similarly to the false positive rates study, the power of all tests studied here is not affected by the latent variable having a mixture of normal distributions and it is lower for small sample sizes. Interestingly, when $\epsilon \sim SN(1)$, the LM(CP) test has the highest power whereas, when $\epsilon \sim SN(3)$, the power is lower for all tests, particularly for LM(S)

in the case of small sample sizes. However, the power, even for $\kappa = 3$, increases with the increase of sample size and number of items. When ϵ is distributed as a skew-normal, the LM(H) test produces non-valid results in some of the simulation scenarios, ranging from 0.2% to 30.2% of the replications and, as in the previous setting, the number of non-valid LM(H) statistics increases with the skewness of the latent variable distribution and decreases as the sample size increases.

The study on the GS(J) test

The GS(J) test is computationally expensive compared to the other tests. Indeed, in each replication of a sample of size n, the Jackknife score covariance matrix given in (14) requires n times the ML-estimates of the parameters. To reduce the time complexity for this method, a faster model estimation is obtained by using the "ltm" R package, which uses a combination of the E-M algorithm and direct maximization. As before, numerical derivatives for the Hessian and cross-product matrix are obtained with the "NumDeriv" R package. We conduct a small-scale simulation to compare the performance of the LM(H), LM(CP), and LM(S) tests with the GS(J) test under no misspecification, misspecification due to local dependence, and misspecification of the latent variable distribution. All models considered here will only have a measurement model and no structural model. We consider the following simulation conditions: number of items ($p = 10$) $\times$ sample size ($n = 200, 500, 1000$) $\times$ test statistic ($LM(H), LM(CP), LM(S), GS(J)$) and 500 replications for each scenario. To study the Type I error/false positive rates, we consider three data generating models (DGM): i) under a correct model specification, data are generated from the 2-PL model (Birnbaum 1968), ii) under local dependence from the model given in equation (15), and iii) under misspecification of the latent variable distribution from the model given in equation (17). To study the power, we set the parameter γ_{1j} equal to 0.5 and 2, on the last item of the three data generating models (2-PL, (15), (17)). For all of them, the covariate x does not affect the latent variable (β=0) and intercepts, factor loadings, and the values of the group variable x are generated as in the Violation of local

independence section. When data are generated from (15), we consider $\sigma_u^2 = 1$ and $LD = 20\%$. For data generated from (17), we assume $\epsilon \sim SN(3)$. We consider the model in equation (16), without the structural model, as the unconstrained model. Under scenario A, γ_{1j} is fixed to 0 under the null hypothesis. Scenario A is used to study the Type I error/false positive rate, because all items in the data generating models are measurement invariant, and to study the power, because a uniform-DIF parameter is introduced on the last item of all data generating models. Table 7 reports the Type I error/false positive rates of the GS(J), LM(H), LM(CP), and LM(S) tests under correct model specification, local dependence, and misspecification of the latent variable distribution, for scenario A.

The GS(J) test and the LM(S) test perform similarly under all conditions. In general, all tests have good performance and only the LM(CP) test shows inflated false positive rates under some conditions.

Table 8 presents the empirical power for the GS(J), LM(H), LM(CP), and LM(S) tests under correct model specification, local dependence, and incorrect distribution of the latent variable, for scenario A.

Under all conditions for small sample size, the power of the GS(J) test is always equal to or lower than the one of the LM(S) test. When the sample size increases, the two tests reach the same power. Similarly to the Type I error/false positive rate study, the performance of the GS(J) test is never superior to that of the other tests. For this reason, and for its high computational cost, we do not use the GS(J) test in the real data analysis.

An application to a real data set

In this section we assess measurement invariance under model misspecification through the LM(H), LM(CP), and LM(S) tests on a real data set, taken from Miller, Swanson, and Newcomb (1984). We select the same sample of observations and items analysed by Duncan (1979). In 1953, in the Detroit Area, the following questions regarding sex role expectations were asked to a sample of 257 women: “Here are some

things that might be done by a boy or a girl. As I read each of these to you, I would like you to tell me if it should be done as a regular task by a boy, by a girl, or by both: (1) Shoveling walks, (2) Washing the car, (3) Dusting furniture, (4) Making beds". Responses of "boy" to items 1 and 2 and "girl" to items 3 and 4 are coded as "0" and refer to traditional answers. Responses of "both" for all items are coded as "1" and refer to "egalitarian" answers. For the same sample of women, in addition to the four binary items, we consider a group variable, that we call "Work", taken from the original data set (Miller et al. 1984). The following question was asked to the sample of mothers "What is your occupation? What kind of business is that in?" The possible responses were the following: "Professional, technical, and kindred workers", "Managers, officials and proprietors, except farm", "Clerical and kindred workers", "Sales workers", "Operatives and kindred workers", "Private household workers, service workers", "Laborers, except farm and mine", and "Not in labor force". We group these responses into two classes:

- Class coded as "0", which includes only answers "Not in labor force". This class includes the group of non-working women ($n_0 = 199$).

- Class coded as "1", which includes all the other responses. This class includes the group of working women ($n_1 = 58$).

The percentages of "egalitarian" answers among the group of non-working women are 31%, 31%, 29% and 42% to items 1-4, respectively. The percentages of "egalitarian" answers among the group of working women are 43%, 29%, 50% and 55% to items 1-4, respectively. Women in the working group give more "egalitarian" answers than women in the non-working group, especially to items 3 and 4. The data set is analysed by Mavridis and Moustaki (2009) and Irincheeva (2011). They show that the classical unidimensional IRT model with the latent variable distributed as a standard normal has a poor fit on this data set. Irincheeva (2011) estimates a semi-nonparametric (SNP) unidimensional IRT model to the data, that allows for more flexibility in the shape of the latent variable distribution, and gives a better fit of the proposed model to the data

compared with the classic unidimensional IRT model. Moreover, the results found by Irincheeva (2011) suggest that the shape of the true latent variable is right skewed or even more complex.

Starting from these results, in this study we consider a unidimensional IRT model for binary data based on the assumption of standard normal latent variable distribution under the null hypothesis, that we know to be misspecified. Measurement invariance on the intercept of each item is tested through $H_0 : \gamma_{1j*} = 0 \ \ vs \ \ H_1 : \gamma_{1j*} \neq 0$, where γ_{1j*} is the effect of the group variable "Work" on the item intercept. Measurement invariance on the item slope of each item is tested through $H_0 : \boldsymbol{\gamma}_{2j*} = 0 \ \ vs \ \ H_1 : \gamma_{2j*} \neq 0$, where γ_{2j*} is the effect of the group variable "Work" on the item slope. Rejecting the null hypothesis implies that the item intercept, or slope, is measurement non-invariant. Due to the small sample size and low number of items, we avoid considering multiple parameter hypothesis testing. The p-values of the tests are computed in two ways, using the asymptotic distribution of the tests under the null hypothesis and bootstrap hypothesis testing (Efron and Tibshirani 1994). As observed in the Simulation study section, under high misspecification of the latent variable distribution, the LM tests do not match their theoretical distributions under the null hypothesis. In particular, the LM(H) and LM(S) tests have the worst performance in terms of power under small sample sizes. The bootstrap hypothesis testing does not depend on the asymptotic distribution of the test statistic under the null hypothesis and can be a good alternative under model misspecification (Lu and Young 2012).

The first step of the bootstrap hypothesis testing procedure is to generate B bootstrap samples, or simulated data sets, indexed by h, that should satisfy the null hypothesis (Efron and Tibshirani 1994). We consider a parametric bootstrap, where the bootstrap samples are generated from a classical unidimensional IRT model with the latent variable distributed as a standard normal and parameter estimates obtained fitting the same model to the original sample of observations. Under the null hypothesis, the group variable "Work" has no effect on the intercept and slope of each item. For this reason, the values of the group variable in each bootstrap sample are

randomly drawn from a Bernoulli variable with success probability estimated on the original sample of observations. The parametric bootstrap can be used even when the model under the null hypothesis is misspecified (Lu and Young 2012). The bootstrap hypothesis testing is composed using the following steps (Efron and Tibshirani 1994):

1. Calculate the statistic $\hat{\tau}$ (the LM(H), LM(CP) and LM(S) tests) in the original sample of observations.

2. Calculate the statistic τ in each bootstrap sample, called τ_h^* .

3. Compute the bootstrap p-value as $\hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{h=1}^{B} I(\tau_h^* > \hat{\tau})$, where I is the indicator function.

4. Reject the null hypothesis if $\hat{p}^*(\hat{\tau}) < \alpha$.

When τ is pivotal, that is its distribution does not depend on unknown parameters, and the number of bootstrap samples B is such that $\alpha(B+1)$ is an integer, the bootstrap hypothesis testing procedure can yield exact test (Dwass 1957). We choose $B = 999$, which is usually a good choice for the number of bootstrap samples to be used in hypothesis testing (MacKinnon 2002).

Table 9 presents the p-values for the LM(H), LM(CP), and LM(S) tests based on their theoretical distributions (TD) under the null hypothesis and on bootstrap hypothesis testing (BH) for measurement invariance on the item intercept and slope.

For all tests, TD and BH do not reject the null hypothesis of intercept and slope invariance for items 1, 2, and 4. This is consistent with the simulation results, in which the false positive rates are less affected than the power of the tests by the misspecification of the latent variable distribution. However, BH and TD disagree for item 3. Interestingly, only the LM(CP) test produces similar results to the BH p-values of the LM(S) test, rejecting the null hypothesis of measurement invariance on the intercept and slope. This is consistent with the simulation results, where the LM(CP) test has the highest power for small sample sizes under misspecification of the latent variable distribution. The bootstrap hypothesis testing procedure for the LM(S) and LM(CP) tests turns out to be a good instrument to make a clearer decision on the

acceptance or rejection of the null hypothesis, especially when these tests show contradictory results. By contrast, the LM(H) test gives negative statistics in the real data set and in a large number of bootstrap replications, as in some simulation scenarios under high misspecification of the latent variable distribution and small sample size. This makes it difficult to interpret results and worsens the performance of the bootstrap hypothesis testing procedure. Indeed, for measurement invariance on the intercept of item 3, the TD and BH p-values of the LM(H) test cannot be computed because the statistic calculated in the real data set is negative. Moreover, in the measurement invariance testing of the slope of item 3, the result of the BH p-value of LM(H) test is not stable because in 11.5% of the bootstrap replications we obtain non-valid statistics that have been excluded from the BH p-value computation.

Discussion

In this work, we evaluated the performance of the LM(H), LM(CP), LM(S), and GS(J) tests to assess measurement invariance under both correct model specification and different types of model misspecification by means of a wide simulation study and in a real data analysis. Moreover, we computed the empirical and asymptotic power of the LM(H), LM(CP), and LM(S) tests, using for the latter the asymptotic distributions of the statistics under the alternative hypothesis.

Under model misspecification, there are some differences between the three tests due to the type and the strength of the model misspecification. Under low local dependence, and when the latent variable is generated from a mixture of normals or from a moderate skew-normal, all tests have good performance in terms of false positive rates and power for large sample sizes. Only the LM(CP) test shows inflated false positive rates in some cases. For this reason, under mild model misspecification, we discourage the use of the LM(CP) test due to its inflated false positive rates. When the misspecification is high, the tests performance deteriorates. Indeed under high local dependence the false positive rates for all tests are seriously inflated while, when the latent variable is highly skewed, with 10 items and for small sample sizes, the LM(H)

and LM(S) tests have very low power. Under high model misspecification, the LM(CP) test has the highest power for small sample sizes. It is worth noting that the LM(S) test, although derived under model misspecification, does not have better performance than the LM(H) test, particularly in terms of power but it always produces valid statistics. Under all types of misspecification considered, we do not find significant differences in the tests' behaviour between the case of measurement invariance on the intercept and that on the intercept and slope, both in single and multiple parameter hypothesis testing.

The simulation study highlights that there are small numerical differences between the asymptotic power, computed through the approximation method for the non-centrality parameter, and the empirical power. However, the results given by the two procedures are coherent and the asymptotic power can be a valid alternative to obtain the power of a test, since it allows us to reduced the time complexity compared to the empirical power.

Concerning the GS(J) test, it is never superior to the other tests and, due to its high computational cost, we do not recommend the use of this test to assess measurement invariance under model misspecification.

Consistently with the simulation results, in the real data analysis the LM(CP) test has the highest power to detect item measurement non-invariance under high misspecification of the latent variable distribution. The bootstrap hypothesis testing procedure turns out to be a good instrument under model misspecification. Indeed, it helps to make a clearer decision on the acceptance or rejection of the null hypothesis when the asymptotic tests provide contradictory results.

For further studies on the performance of the LM tests under model misspecification, different types of estimation methods could be considered. Moreover, we found that when data are generated assuming a skew-normal distribution for the latent variable, parameter estimates are seriously biased with respect to the true parameters' values. Further research should be devoted to exploring misspecified models where the parameter estimates are consistent with respect to the true parameter values.

In these cases, the LM tests should have a better performance.

References

Agresti, A. (2002). *Categorical data analysis*. John Wiley & Sons.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

Bera, A. K., Doğan, O., Bilias, Y., Yoon, M. J., Taşpınar, S., et al. (2020). Adjustments of Rao's score test for distributional and local parametric misspecifications. *Journal of Econometric Methods*, *9*(1), 1–29.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

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

Boos, D. D. (1992). On generalized score tests. *The American Statistician*, *46*(4), 327–333.

Boos, D. D., & Stefanski, L. A. (2013). Hypothesis tests under misspecification and relaxed assumptions. In *Essential statistical inference: Theory and methods* (pp. 339–359). New York, NY: Springer New York.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

Duncan, O. D. (1979). Indicators of sex typing: Traditional and egalitarian, situational and ideological responses. *American Journal of Sociology*, *85*(2), 251–260.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 181–187.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Engle, R. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (1st ed., Vol. 2, p. 775-826). Elsevier.

Falk, C. F., & Monroe, S. (2018). On Lagrange multiplier tests in multidimensional

item response theory: Information matrices and model misspecification. *Educational and Psychological Measurement*, *78*(4), 653–678.

Fox, J., & Glas, C. A. (2005). Bayesian modification indices for IRT models. *Statistica Neerlandica*, *59*(1), 95–106.

Glas, C. A. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *98*(8), 647–667.

Glas, C. A. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika*, *64*(3), 273–294.

Glas, C. A., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106.

Godambe, V., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review/Revue Internationale de Statistique*, 127–138.

Green, S. B., Thompson, M. S., & Babyak, M. A. (1998). A Monte Carlo investigation of methods for controlling Type I errors with specification searches in structural equation modeling. *Multivariate Behavioral Research*, *33*(3), 365–383.

Guastadisegni, L., Cagnone, S., Moustaki, I., & Vasdekis, V. (in press). The asymptotic power of the lagrange multiplier tests for misspecified IRT models. In *Quantitative psychology: The 85th annual meeting of the psychometric society virtual, 2020.* Springer.

Gudicha, D. W., Schmittmann, V. D., & Vermunt, J. K. (2017). Statistical power of likelihood ratio and Wald tests in latent class models with covariates. *Behavior Research Methods*, *49*(5), 1824–1837.

Irincheeva, I. (2011). *Generalized linear latent variable models with flexible distributions* (Unpublished doctoral dissertation). University of Geneva.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple

indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639.

Kim, D., De Ayala, R., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement*, *35*(6), 447–471.

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254-274.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*, *36*(8), 670–688.

Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 496–513.

Lu, H. K., & Young, G. A. (2012). Parametric bootstrap under model mis-specification. *Computational Statistics & Data Analysis*, *56*(8), 2410–2420.

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique*, *35*(4), 615–645.

Mavridis, D., & Moustaki, I. (2009). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, *18*(4), 1016–1034.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105–118.

Mellenbergh, G. J. (1983). Conditional item bias methods. In *Human assessment and cultural factors* (pp. 293–302). Springer.

Miller, D., Swanson, G. E., & Newcomb, T. M. (1984). *Detroit area study, 1953: Child training patterns among urban families and attitudes and perceptions of consensus of group members*. Inter-university Consortium for Political and Social Research.

Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class

models. *Advances in Data Analysis and Classification*, *7*(3), 267–279.

Ranger, J., & Kuhn, J.-T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement*, *49*(3), 247–268.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical proceedings of the Cambridge philosophical society* (Vol. 44, pp. 50–57).

Rao, J., Scott, A. J., & Skinner, C. J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 1059–1070.

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, 331–360.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*(1), 131–151.

Shao, J. (1992). Jackknifing in generalized linear models. *Annals of the Institute of Statistical Mathematics*, *44*(4), 673–686.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.

van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.

Table 1

False positive rates of the LM(H), LM(CP), and LM(S) tests under scenarios A, B and C, $p = 10$, $n = 200, 500, 1000, 5000$

				$\sigma_u^2 = 0.25$			$\sigma_u^2 = 1$			$\sigma_u^2 = 2.25$		
SC	p	LD	n	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)
A	10	20%	200	0.05	0.066	0.052	0.044	0.066	0.034	0.044	**0.082**	0.052
			500	**0.072**	**0.08**	**0.074**	**0.072**	**0.084**	**0.078**	**0.086**	**0.104**	**0.088**
			1000	0.064	**0.076**	**0.07**	0.05	0.054	0.052	**0.09**	**0.112**	**0.104**
			5000	0.046	0.05	0.048	**0.092**	**0.098**	**0.094**	**0.23**	**0.246**	**0.246**
		50%	200	0.042	**0.078**	0.044	0.044	**0.08**	0.052	**0.092**	**0.168**	**0.112**
			500	**0.072**	**0.082**	**0.074**	**0.116**	**0.148**	**0.134**	**0.256**	**0.298**	**0.282**
			1000	**0.076**	**0.08**	**0.072**	**0.152**	**0.184**	**0.17**	**0.412**	**0.458**	**0.446**
	20	20%	200	0.04	**0.094**	0.05	0.056	**0.09**	0.056	0.06	**0.118**	0.068
			500	0.044	0.06	0.048	0.058	**0.078**	**0.07**	**0.092**	**0.108**	**0.096**
			1000	0.046	0.054	0.052	**0.076**	**0.088**	**0.078**	**0.152**	**0.174**	**0.162**
		50%	200	0.052	**0.11**	0.06	**0.074**	**0.13**	**0.088**	**0.15**	**0.242**	**0.178**
			500	0.052	**0.076**	0.058	**0.132**	**0.168**	**0.148**	**0.334**	**0.388**	**0.358**
			1000	0.054	**0.07**	0.064	**0.188**	**0.224**	**0.212**	**0.58**	**0.622**	**0.604**
B	10	20%	200	**0.1**	**0.122**	0.052	**0.092**	**0.106**	0.036	**0.074**	**0.112**	**0.044**
			500	0.062	**0.07**	0.042	0.066	**0.082**	0.054	**0.076**	**0.088**	**0.058**
			1000	0.064	0.064	0.048	0.046	0.066	0.05	**0.094**	**0.094**	**0.086**
		50%	200	0.062	**0.124**	0.036	**0.11**	**0.190**	**0.078**	**0.394**	**0.386**	**0.148**
			500	0.05	**0.092**	0.044	**0.236**	**0.298**	**0.226**	**0.796**	**0.71**	**0.61**
			1000	0.068	**0.096**	**0.08**	**0.492**	**0.456**	**0.426**	**0.978**	**0.954**	**0.942**
	20	20%	200	**0.03**	**0.162**	0.032	0.06	**0.194**	0.05	**0.082**	**0.208**	0.068
			500	0.048	**0.074**	0.048	0.06	**0.09**	0.056	**0.144**	**0.114**	**0.08**
			1000	0.04	0.054	0.046	**0.082**	**0.084**	0.066	**0.246**	**0.16**	**0.132**
		50%	200	0.036	**0.178**	0.04	**0.11**	**0.26**	**0.098**	**0.288**	**0.442**	**0.214**
			500	0.058	**0.096**	0.066	**0.206**	**0.244**	**0.18**	**0.648**	**0.608**	**0.518**
			1000	0.064	**0.096**	**0.072**	**0.418**	**0.384**	**0.34**	**0.946**	**0.916**	**0.886**
C	10	20%	200	0.06	**0.104**	0.04	0.058	**0.094**	0.046	0.066	**0.112**	0.046
			500	0.068	**0.092**	0.068	0.056	**0.08**	0.054	0.06	**0.118**	**0.08**
			1000	0.064	0.068	0.056	0.042	0.06	0.052	**0.086**	**0.128**	**0.112**
		50%	200	0.062	**0.102**	0.036	0.056	**0.122**	0.05	**0.094**	**0.214**	**0.086**
			500	0.062	**0.086**	0.062	**0.084**	**0.14**	**0.098**	**0.2**	**0.278**	**0.22**
			1000	0.058	**0.08**	0.068	**0.11**	**0.154**	**0.142**	**0.34**	**0.398**	**0.364**
	20	20%	200	0.056	**0.156**	0.052	0.056	**0.138**	0.06	0.062	**0.172**	0.066
			500	**0.072**	**0.092**	**0.07**	0.05	**0.098**	**0.074**	0.06	**0.11**	**0.07**
			1000	0.048	0.068	0.052	0.06	**0.09**	**0.072**	**0.122**	**0.17**	**0.146**
		50%	200	0.064	**0.16**	0.058	0.052	**0.17**	0.068	**0.124**	**0.286**	**0.146**
			500	0.064	**0.086**	0.062	**0.112**	**0.172**	**0.112**	**0.256**	**0.36**	**0.284**
			1000	0.064	**0.078**	**0.07**	**0.132**	**0.172**	**0.156**	**0.494**	**0.538**	**0.52**

Note 1: Values in boldface indicate that the nominal level α is not included in their confidence interval

Table 2

Empirical power (EP) and asymptotic power (AP) of the LM(H), LM(CP), and LM(S) tests under scenario D, $p = 10, 20$*,* $n = 200, 500, 1000, 5000$

					$\sigma_u^2 = 0.25$			$\sigma_u^2 = 1$			$\sigma_u^2 = 2.25$		
SC	p	LD	n		LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)
D	10	20%	200	EP	0.308	0.398	0.32	0.38	0.452	0.388	0.484	0.55	0.494
				AP	0.459	0.506	0.485	0.473	0.514	0.493	0.543	0.584	0.562
			500	EP	0.702	0.724	0.71	0.776	0.806	0.798	0.864	0.878	0.872
				AP	0.836	0.877	0.859	0.849	0.884	0.867	0.905	0.930	0.917
			1000	EP	0.936	0.942	0.938	0.97	0.974	0.974	0.994	0.994	0.994
				AP	0.985	0.993	0.990	0.988	0.994	0.991	0.996	0.998	0.997
			5000	EP	1	1	1	1	1	1	1	1	1
				AP	1	1	1	1	1	1	1	1	1
		50%	200	EP	0.324	0.44	0.356	0.49	0.57	0.516	0.637	0.706	0.624
				AP	0.497	0.552	0.527	0.586	0.649	0.621	0.723	0.777	0.739
			500	EP	0.752	0.774	0.758	0.888	0.898	0.89	0.956	0.96	0.956
				AP	0.870	0.911	0.893	0.931	0.959	0.948	0.981	0.990	0.984
			1000	EP	0.952	0.956	0.952	0.992	0.994	0.992	1	1	1
				AP	0.992	0.997	0.995	0.998	0.999	0.999	1	1	1
	20	20%	200	EP	0.382	0.528	0.392	0.484	0.606	0.484	0.574	0.66	0.582
				AP	0.473	0.506	0.492	0.523	0.557	0.542	0.570	0.603	0.588
			500	EP	0.824	0.858	0.83	0.886	0.910	0.889	0.94	0.946	0.936
				AP	0.849	0.877	0.866	0.891	0.914	0.904	0.922	0.939	0.932
			1000	EP	0.982	0.986	0.982	0.994	0.994	0.994	1	1	1
				AP	0.988	0.993	0.991	0.995	0.997	0.996	0.997	0.998	0.998
		50%	200	EP	0.416	0.558	0.42	0.59	0.68	0.592	0.74	0.832	0.742
				AP	0.497	0.531	0.517	0.624	0.668	0.649	0.752	0.794	0.772
			500	EP	0.844	0.866	0.846	0.962	0.97	0.964	0.992	0.994	0.992
				AP	0.870	0.896	0.886	0.949	0.966	0.959	0.986	0.992	0.989
			1000	EP	0.992	0.994	0.994	1	1	1	1	1	1
				AP	0.992	0.995	0.994	0.999	1	0.999	1	1	1

Table 3

Empirical power of the LM(H), LM(CP), and LM(S) tests under scenarios E and F, $p = 10, 20$, $n = 200, 500, 1000, 5000$

				$\sigma_u^2 = 0.25$			$\sigma_u^2 = 1$			$\sigma_u^2 = 2.25$		
SC	p	LD	n	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)
E	10	20%	200	0.449	0.58	0.37	0.502	0.606	0.412	0.538	0.624	0.432
			500	0.9	0.926	0.902	0.934	0.948	0.928	0.966	0.974	0.958
			1000	0.998	1	1	0.996	0.998	0.998	0.998	1	0.998
		50%	200	0.518	0.606	0.364	0.730	0.716	0.372	0.858	0.779	0.3
			500	0.948	0.954	0.926	0.994	0.984	0.968	0.998	0.998	0.978
			1000	1	1	0.998	1	1	1	1	1	1
	20	20%	200	0.742	0.876	0.722	0.802	0.856	0.692	0.834	0.866	0.722
			500	0.994	0.996	0.994	1	0.998	0.994	1	1	0.994
			1000	1	1	1	1	1	1	1	1	1
		50%	200	0.814	0.906	0.966	0.9	0.934	0.818	0.966	0.962	0.894
			500	1	1	0.998	1	1	1	1	1	1
			1000	1	1	1	1	1	1	1	1	1
F	10	20%	200	0.660	0.632	0.416	0.674	0.662	0.486	0.743	0.758	0.598
			500	0.957	0.946	0.898	0.978	0.976	0.944	0.992	0.99	0.98
			1000	0.998	0.998	0.998	1	1	1	1	1	1
		50%	200	0.637	0.61	0.388	0.641	0.636	0.398	0.662	0.617	0.381
			500	0.945	0.932	0.902	0.951	0.94	0.91	0.940	0.926	0.894
			1000	0.998	0.998	0.996	1	0.998	0.998	1	1	1
	20	20%	200	0.807	0.848	0.666	0.860	0.888	0.756	0.896	0.91	0.802
			500	0.992	0.996	0.982	0.996	0.996	0.996	1	1	0.998
			1000	1	1	1	1	1	1	1	1	1
		50%	200	0.803	0.844	0.664	0.852	0.872	0.696	0.823	0.862	0.682
			500	0.992	0.996	0.984	0.996	0.996	0.992	0.991	0.996	0.99
			1000	1	1	1	1	1	1	1	1	1

Table 4

False positive rates of the LM(H), LM(CP), and LM(S) tests under scenarios A, B and C, $p = 10, 20$, $n = 200, 500, 1000$

			$\epsilon \sim 0.3N(-1.5, 0.2) + 0.7N(1, 0.4)$			$\epsilon \sim SN(1)$			$\epsilon \sim SN(3)$		
SC	p	n	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)
A	10	200	0.048	0.066	0.042	0.046	**0.076**	**0.024**	**0.089**	**0.132**	**0.008**
		500	0.046	0.052	0.04	0.05	0.066	0.042	**0.076**	**0.07**	**0.022**
		1000	0.048	0.052	0.05	0.06	0.062	0.056	0.06	0.058	0.042
	20	200	0.054	**0.082**	0.056	0.054	**0.116**	0.044	0.06	**0.112**	**0.026**
		500	0.05	0.058	0.05	0.054	0.066	0.058	0.056	**0.07**	0.044
		1000	0.042	0.04	0.038	0.052	**0.07**	0.066	0.054	0.06	0.054
B	10	200	0.06	**0.10**	0.046	**0.134**	**0.156**	**0.016**	**0.198**	**0.242**	**0.002**
		500	0.058	0.066	0.048	**0.112**	**0.09**	0.032	**0.195**	**0.082**	**0.004**
		1000	0.066	0.066	0.058	**0.086**	0.06	0.042	**0.196**	0.066	**0.002**
	20	200	0.058	**0.140**	0.042	0.066	**0.222**	0.04	**0.119**	**0.293**	**0.002**
		500	0.044	0.064	0.034	0.056	**0.102**	0.044	0.066	**0.114**	**0.016**
		1000	0.064	**0.076**	0.054	0.042	0.064	0.05	**0.072**	**0.09**	0.042
C	10	200	**0.07**	**0.118**	0.048	0.065	**0.164**	**0.026**	**0.133**	**0.216**	**0.012**
		500	0.066	**0.072**	0.036	0.05	**0.078**	0.042	**0.075**	**0.092**	0.032
		1000	0.062	0.068	0.056	0.066	0.068	0.052	**0.076**	**0.084**	**0.026**
	20	200	**0.076**	**0.154**	0.046	0.062	**0.218**	0.042	**0.087**	**0.235**	**0.02**
		500	0.05	**0.094**	0.044	0.044	**0.084**	0.046	0.046	**0.09**	**0.03**
		1000	0.068	**0.084**	0.056	0.044	0.064	0.042	**0.07**	**0.098**	0.048

Note 1: Values in boldface indicate that the nominal level α is not included in their confidence interval

Table 5

Empirical power (EP) and asymptotic power (AP) of the LM(H), LM(CP), and LM(S) tests under scenario D, $p = 10, 20$, $n = 200, 500, 1000$

				$\epsilon \sim 0.3N(-1.5, 0.2) + 0.7N(1, 0.4)$			$\epsilon \sim SN(1)$			$\epsilon \sim SN(3)$		
SC	p	n		LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)
D	10	200	EP	0.316	0.396	0.324	0.195	0.28	0.15	0.129	0.186	0.03
			AP	0.425	0.459	0.443	0.307	0.326	0.301	0.226	0.208	0.170
		500	EP	0.684	0.71	0.7	0.424	0.462	0.406	0.235	0.244	0.094
			AP	0.772	0.799	0.835	0.632	0.664	0.623	0.480	0.440	0.354
		1000	EP	0.95	0.958	0.952	0.748	0.762	0.75	0.406	0.402	0.328
			AP	0.977	0.986	0.982	0.902	0.921	0.895	0.771	0.725	0.611
	20	200	EP	0.38	0.488	0.382	0.292	0.414	0.282	0.197	0.299	0.092
			AP	0.385	0.400	0.392	0.397	0.421	0.391	0.232	0.237	0.218
		500	EP	0.76	0.804	0.768	0.596	0.64	0.586	0.406	0.464	0.354
			AP	0.751	0.770	0.759	0.766	0.794	0.759	0.492	0.502	0.461
		1000	EP	0.98	0.98	0.978	0.902	0.906	0.898	0.662	0.692	0.644
			AP	0.961	0.968	0.965	0.967	0.976	0.964	0.783	0.794	0.749

Table 6

Empirical power of the LM(H), LM(CP), and LM(S) tests under scenarios E and F, $p = 10, 20$*,* $n = 200, 500, 1000$

			$\epsilon \sim 0.3N(-1.5, 0.2) + 0.7N(1, 0.4)$			$\epsilon \sim SN(1)$			$\epsilon \sim SN(3)$		
SC	p	n	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)	LM(H)	LM(CP)	LM(S)
E	10	200	0.516	0.614	0.402	0.218	0.446	0.124	0.100	0.313	0.02
		500	0.926	0.93	0.91	0.627	0.756	0.632	0.347	0.408	0.09
		1000	0.998	0.998	0.998	0.946	0.972	0.962	0.642	0.7	0.312
	20	200	0.674	0.853	0.646	0.524	0.782	0.456	0.385	0.642	0.076
		500	0.992	0.996	0.99	0.946	0.968	0.946	0.739	0.81	0.488
		1000	1	1	1	1	1	1	0.974	0.98	0.954
F	10	200	0.588	0.547	0.318	0.356	0.484	0.188	0.223	0.462	0.158
		500	0.916	0.89	0.838	0.834	0.844	0.722	0.585	0.772	0.532
		1000	0.99	0.988	0.988	0.974	0.982	0.972	0.867	0.966	0.882
	20	200	0.449	0.48	0.174	0.713	0.787	0.52	0.608	0.783	0.434
		500	0.826	0.784	0.7	0.988	0.986	0.97	0.921	0.984	0.952
		1000	0.978	0.97	0.952	1	1	1	0.958	1	1

Table 7

Type I error/ false positive rate of the GS(J), LM(H), LM(CP), and LM(S) tests under scenario A, $p = 10$*,* $n = 200, 500, 1000$

Data generating model	*SC*	p	n	GS(J)	LM(H)	LM(CP)	LM(S)
2-PL	A	10	200	0.042	0.048	0.064	0.046
			500	0.06	0.06	**0.072**	0.06
			1000	0.062	0.062	0.062	0.062
(15)	A	10	200	0.034	0.042	0.054	0.034
			500	0.056	0.058	0.064	0.056
			1000	0.056	0.058	0.064	0.058
(17)	A	10	200	0.036	0.044	**0.072**	0.036
			500	0.044	0.048	0.058	0.044
			1000	0.048	0.052	0.056	0.048

Note 1: Values in boldface indicate that the nominal level α is not included in their confidence interval

Table 8

Empirical power of the GS(J), LM(H), LM(CP), and LM(S) tests under scenario A, $p = 10$*,* $n = 200, 500, 1000$

Data generating model	*SC*	p	γ_{1j}	n	GS(J)	LM(H)	LM(CP)	LM(S)
2-PL	A	10	0.5	200	0.23	0.292	0.296	0.238
				500	0.488	0.53	0.52	0.494
				1000	0.754	0.778	0.772	0.758
			2	200	0.962	0.98	0.982	0.962
				500	1	1	1	1
				1000	1	1	1	1
(15)	A	10	0.5	200	0.176	0.236	0.234	0.186
				500	0.394	0.434	0.422	0.396
				1000	0.67	0.686	0.676	0.67
			2	200	0.956	0.978	0.978	0.962
				500	1	1	1	1
				1000	1	1	1	1
(17)	A	10	0.5	200	0.11	0.200	0.196	0.13
				500	0.344	0.414	0.392	0.344
				1000	0.62	0.678	0.634	0.622
			2	200	0.634	0.893	0.903	0.732
				500	0.996	1	0.998	0.996
				1000	1	1	1	1

Table 9

Theoretical distributions (TD) and bootstrap hypothesis testing (BH) p-values of the LM(H), LM(CP), and LM(S) tests for measurement invariance on the item intercept and slope

Parameter tested	Item	Method	LM(H)	LM(CP)	LM(S)
γ_{1j*}	1	TD	0.387	0.390	0.391
		BH	0.397	0.404	0.398
	2	TD	0.107	0.082	0.097
		BH	0.114	0.102	0.105
	3	TD	-	**0.014**	0.059
		BH	-	**0.023**	**0.020**
	4	TD	0.78	0.795	0.801
		BH	0.800	0.811	0.811
γ_{2j*}	1	TD	0.399	0.351	0.353
		BH	0.393	0.346	0.337
	2	TD	0.116	0.112	0.131
		BH	0.124	0.118	0.114
	3	TD	**0.048**	**0.038**	0.098
		BH	0.101	**0.049**	**0.031**
	4	TD	0.050	0.118	0.223
		BH	0.083	0.163	0.172

Note 1: Values in boldface indicate p-values less than the nominal level α