Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo 34

Settore Concorsuale: 13/D1 - STATISTICA

Settore Scientifico Disciplinare: SECS-S/01 - STATISTICA

ASSESSING THE FIT OF UNIDIMENSIONAL IRT MODELS FOR BINARY DATA UNDER MODEL MISSPECIFICATION

Presentata da: Lucia Guastadisegni

Coordinatore Dottorato

Monica Chiogna

Supervisore

Silvia Cagnone

Co-Supervisori

Irini Moustaki

Vassilis Vasdekis

Esame finale anno 2022

Abstract

Model misspecification affects the classical test statistics used to assess the fit of the Item Response Theory (IRT) models. Robust tests have been derived under model misspecification, as the Generalized Lagrange Multiplier and Hausman tests, but their use has not been largely explored in the IRT framework.

In the first part of the thesis, we introduce the Generalized Lagrange Multiplier test to detect differential item response functioning in IRT models for binary data under model misspecification. By means of a simulation study and a real data analysis, we compare its performance with the classical Lagrange Multiplier test, computed using the Hessian and the cross-product matrix, and the Generalized Jackknife Score test. The power of these tests is computed empirically and asymptotically. The misspecifications considered are local dependence among items and non-normal distribution of the latent variable. The results highlight that, under mild model misspecification, all tests have good performance while, under strong model misspecification, the performance of the tests deteriorates. None of the tests considered show an overall superior performance than the others.

In the second part of the thesis, we extend the Generalized Hausman test to detect non-normality of the latent variable distribution. To build the test, we consider a seminonparametric-IRT model, that assumes a more flexible latent variable distribution. By means of a simulation study and two real applications, we compare the performance of the Generalized Hausman test with the M_2 limited information goodness-of-fit test and the Likelihood-Ratio test. Additionally, the information criteria are computed. The Generalized Hausman test has a better performance than the Likelihood-Ratio test in terms of Type I error rates and the M_2 test in terms of power. The performance of the Generalized Hausman test and the information criteria deteriorates when the sample size is small and with a few items.

To my parents

Acknowledgements

First of all, I would like to express my heartfelt gratitude to my supervisor Prof. Silvia Cagnone, for her dedicated support and great encouragement throughout this research project. Her constructive feedback, expertise and guidance over the past three years have been fundamental in overcoming the difficulties encountered along the way and completing the thesis.

A special acknowledgement goes to Prof. Irini Moustaki, for her advice during my visiting period abroad at the London School of Economics and Political Science. Her stimulating and inspiring ideas, experience and motivation have been essential to develop the thesis project.

My sincerest thanks go also to Prof. Vassilis Vasdekis, that helped and encouraged me during this working process. The thesis has been enriched by his knowledge in this field, valuable ideas and suggestions.

I acknowledge the financial support from the ERASMUS+ program for my stay in London.

Finally, a special thought goes to my family and friends, for their unconditional support during these challenging yet fulfilling years of PhD.

Contents

List of Tables

Chapter 1

Introduction

In many fields of research, when multivariate data are analyzed, the variables of interest can be theoretical constructs, such as intelligence, quality of life, business confidence, that cannot be observed and measured directly. Latent variable models are statistical models that, through the analysis of the association or correlation among the observed variables, extract information about the constructs of interest, represented in the model by the latent variables. The latent and manifest variables can be either continuous or discrete. Latent trait or, equivalently, Item Response Theory (IRT) models correspond to the case of binary or polytomous outcomes and continuous latent variables (Bartholomew et al., 2011). Binary outcomes are very common in the fields of social, psychological and educational sciences, where IRT models are used to measure fundamental attitudes or abilities.

For these models, the tests commonly used for hypothesis testing are the standard Likelihood-Ratio, Wald, and Lagrange Multiplier or score test statistics (Cox and Hinkley, 1979). The latter has been widely used to detect different types of IRT model violations (Glas, 1998, Fox and Glas, 2005, Glas, 1999, Glas and Falcón, 2003, Kim et al., 2011, Liu and Thissen, 2012, Liu and Maydeu-Olivares, 2013, Liu and Thissen, 2014, van der Linden and Glas, 2010, Oberski et al., 2013, Ranger and Kuhn, 2012) and can be preferred to the Likelihood-Ratio and Wald test because it is computationally less intensive.

The overall goodness-of-fit of an IRT model is usually assessed through the Pearson's chi-square and the Likelihood-Ratio test (Agresti, 2002). However, these tests can be affected by the problem of sparse data (Koehler and Larntz, 1980), that occur very frequently with binary items. To overcome this problem, Maydeu-Olivares and Joe (2005) propose the limited information test statistic M_2, that is based on a quadratic form in marginal residuals of order one and two.

However, under model misspecification, the IRT model fit is an open issue. When hypothesis testing is performed on a misspecified model, the Likelihood-Ratio, the Wald, and the Lagrange Multiplier tests do not have the expected distribution under the null hypothesis. Also the overall goodness-of-fit tests may be affected by model misspecification. For example, the M_2 test lacks of power to reject the fitted model when the item characteristic curve is misspecified on some items (Ranger and Much, 2020).

White (1982) studied the problem of model misspecification when maximum-likelihood based techniques are used. He developed the generalized Wald, Likelihood-Ratio and Lagrange Multiplier test statistics, which allow to draw robust inference when the model is misspecified. He derived also a generalized version of the Hausman test (Hausman, 1978), that is a specification test that compares two types of estimators, the first one that is consistent only under correct model specification, the second one under both correct and misspecified models.

In the IRT context, as far as we know, only the Generalized Lagrange Multiplier test has been studied by Falk and Monroe (2018), to test a single omitted cross-loading under model misspecification.

The first objective of the thesis is to evaluate the performance of the Generalized Lagrange Multiplier test considering a more general framework than Falk and Monroe (2018). In more details, we propose to use the Generalized Lagrange Multiplier test to detect differential item functioning under violations of two fundamental IRT model assumptions, that is local dependence and non-normality of the latent variable distribution. By means of a simulation study and in a real data application, we compare the classical Lagrange Multiplier test, computed using the Hessian and the cross-product matrix, with the Generalized Lagrange Multiplier test. For the latter also a second version is considered, the Generalized Jackknife Score test (J. Rao et al., 1998), where the covariance matrix of the score is computed with the Jackknife method. We also provide two procedures to compute the power of the tests using their asymptotic distribution under the alternative hypothesis.

Motivated by the simulation results that highlight that as the misspecification increases the performance of all the tests analyzed deteriorates, we then focus on the case of misspecification of the latent variable distribution. We consider the semi-nonparametric (SNP) distribution for the latent variable, introduced by Gallant and Tauchen (1989), and studied in the IRT framework by Woods and Lin (2009). This approach allows for a more flexible smooth density of the latent variable.

The second objective of the thesis is to extend the Generalized Hausman test to detect non-normality of the latent variable distribution using the SNP-IRT model. To build the test, we compare the estimators resulting from the classic IRT model that assumes the normality of the latent variable with those resulting from the SNP-IRT model, that better captures the non-normality of the latent variable distribution. Similarly to Ranger and Much (2020), who compare the performance of the classical Hausman test with the M_2 test to detect local dependence and an incorrect specification of the item characteristic curve, we compare the Generalized Hausman test with the M_2 test to detect non-normality of the latent variable distribution, by means of a simulation study and in real data applications. Since the classic IRT and the SNP-IRT models are nested, we also compare the performance of the Generalized Hausman test with the Likelihood-Ratio test.

The thesis is organized as follows. Chapter 2 describes the IRT model for binary data, with a covariate included in the model. We present the different versions of the Lagrange Multiplier test and the Generalized Lagrange Multiplier test, with the related procedures to compute the asymptotic power of tests. Chapter 2 provides also a wide simulation study, that allows to evaluate the performance of the tests considered under model misspecification, in terms of false positive rates, empirical and asymptotic power. The simulation schemes are designed to study the tests performance under different levels of model misspecification, from a mild to a high level, and to test single and multiple parameter for measurement invariance. A real data application is included in this Chapter. The contents of Chapter 2 have been published in Guastadisegni et al. (2021) and Guastadisegni et al. (2022). In Chapter

3 we first provide the description of the SNP-IRT model for binary data. Then we review the M_2 test and information criteria and we present the Generalized Hausman test to detect non-normality of the latent variable distribution. A simulation study is carried out to evaluate the performance of the Generalized Hausman test in terms of Type I error rates and power to detect non-normality of the latent variable distribution. We compare it with the M_2 and the Likelihood-Ratio tests. Information criteria are also computed in all the simulation scenarios. To conclude, the use of the Generalized Hausman test is illustrated in two real data examples.

.

Chapter 2

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

2.1 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 associated 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) a correct specification of the 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 et al., 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) are typically used for hypothesis testing 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 et al., 2011, Liu and Thissen, 2012, Liu and Maydeu-Olivares, 2013, Liu and Thissen, 2014, van der Linden and Glas, 2010, Oberski et al., 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 with many items. 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, 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 many items 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). Different versions of the LM test are also derived under model misspecification. 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 et al., 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). Falk and Monroe (2018) compare the performance of the LM(S), LM(CP), and LM(H) tests for a single omitted cross-loading. In this thesis, 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 latter 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 Chapter 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 et al., 2017). The non-centrality parameter is approximated using the procedure derived by Gudicha et al. (2017) for the Wald and Likelihood-Ratio tests. We extend this method to the LM tests and under model misspecification. Moreover, we propose a second procedure to compute the asymptotic power of the LM tests.

Through some extensive simulation studies, we compare the performance of the different versions of the LM tests in terms of Type I error rate, false positive rate, 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 Chapter 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 procedures to estimate the asymptotic power for the LM(H), LM(CP), and LM(S) tests. Next, we present some Monte Carlo simulation studies and the results from the real data analysis. Finally, some concluding remarks are presented and discussed.

2.2 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:

$$\begin{gathered} P(y_{ij} = 1 | z_i, x_i) = \pi_{ij}(z_i, x_i) = \frac{\exp\left(\alpha_{0j} + \alpha_{1j} z_i + \gamma_{1j} x_i + \gamma_{2j} x_i z_i\right)}{1 + \exp\left(\alpha_{0j} + \alpha_{1j} z_i + \gamma_{1j} x_i + \gamma_{2j} x_i z_i\right)} \\ z_i = \beta x_i + \epsilon_i \qquad \epsilon \sim N(0,1) \end{gathered} \tag{2.1}$$

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:

$$l(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\mathbf{y}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \int \prod_{j=1}^{p} \pi_{ij}(z_i, x_i)^{y_{ij}} (1 - \pi_{ij}(z_i, x_i))^{1-y_{ij}} \phi(z_i \mid x_i) dz_i, \tag{2.2}$$

where $\boldsymbol{\theta}$ is the vector of the unknown parameters and the model assumes conditional/local independence among the items. Equation (2.2) 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 $(\gamma_{1j},\gamma_{2j})$ 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.

2.3 Lagrange Multiplier tests

2.3.1 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:

$$H_0 : \boldsymbol{\theta}'_{02} = \mathbf{c} \quad vs \quad H_1 : \boldsymbol{\theta}'_{02} \neq \mathbf{c}, \tag{2.3}$$

where $\mathbf{c}$ is a vector of constants. Commonly, in real data analysis, $\mathbf{c}$ is the null vector and rejecting the null hypothesis reveals the presence of DIF on the intercept or intercept and slope of one or more items. For this reason, in sections 2.4, 2.6 and 2.7, we consider null hypothesis where $\mathbf{c}$ is the null vector. However, to assess the performance of the LM test under different conditions, in section 2.5, for the power analysis we consider scenarios where $\mathbf{c}$ is a vector of constants different from 0 and DIF is not present in the data generating models.

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

$$LM = S(\tilde{\boldsymbol{\theta}})' A_n(\tilde{\boldsymbol{\theta}})^{-1} S(\tilde{\boldsymbol{\theta}}), \tag{2.4}$$

where $\tilde{\boldsymbol{\theta}}' = (\tilde{\boldsymbol{\theta}}'_1, \mathbf{c})$ denotes the restricted maximum likelihood estimates of the parameters $\boldsymbol{\theta}$, $S(\tilde{\boldsymbol{\theta}}) = \frac{\partial l(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ 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 $\tilde{\boldsymbol{\theta}}_1$ is $\mathbf{0}$, the LM statistic given in (2.4) is reduced to

$$LM = S_2(\tilde{\boldsymbol{\theta}}) A_n^{22}(\tilde{\boldsymbol{\theta}})^{-1} S_2(\tilde{\boldsymbol{\theta}}), \tag{2.5}$$

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)

$$A_n^{22} = A_{n22} - A_{n21} A_{n11}^{-1} A_{n12}, \tag{2.6}$$

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

$$\hat{A}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2 l(\mathbf{y}_i, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \tag{2.7}$$

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

$$\hat{B}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial l(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial l(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} \tag{2.8}$$

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)

$$\lambda = \frac{1}{n}\boldsymbol{\xi}' A_n^{22}(\boldsymbol{\theta_0})\boldsymbol{\xi} \tag{2.9}$$

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

Approximation procedures for the asymptotic power

The asymptotic distribution of the LM test under the alternative hypothesis as a non-central chi-square with non-centrality parameter (2.9) 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 (2.3) (Gudicha et al., 2017). We present here two different approximation procedures for the computation of the non-centrality parameter.

The first method extends the approximation procedure for the asymptotic power derived by Gudicha et al. (2017) for the Likelihood-Ratio and the Wald tests to the LM(H)/LM(CP) test. It can be summarized in the following steps:

1. From the model defined under the alternative hypothesis, create a large data data set (e.g. $N = 10000$ observations).
2. Fit the model under H_0 to the data.
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).

We propose a second method, that is also is also based on some of the steps of the procedure proposed by Gudicha et al. (2017), but the non-centrality parameter is computed according to formula (2.9). The procedure can be summarized as follows:

1. From the model defined under the alternative hypothesis, create a large data data set (e.g. $N = 10000$ observations).
2. Fit the model under H_0 to the data.
3. Compute $\boldsymbol{\xi} = \sqrt{N}(\boldsymbol{\theta}_{02} - \mathbf{c})$, where $\boldsymbol{\theta}_{02}$ is the vector of the data generating values (values under H_1) of the constrained parameters and $\mathbf{c}$ is the vector of constants under the null hypothesis (Reiser, 2008).
4. Compute the non-centrality parameter of the LM(H)/LM(CP) test according to formula (2.9) where $A_n^{22}(\boldsymbol{\theta}^0)$ can be consistently estimated by the corresponding matrix $\hat{A}_n^{22}/\hat{B}_n^{22}$, evaluated at $\tilde{\boldsymbol{\theta}}$.

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

6. 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 is computed as in the first procedure, but using the non-centrality parameter computed at point 6.

2.3.2 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 (2.7) and (2.8) 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 section 2.3.1. The hypotheses in (2.3) can be formalized as follows

$$H_0 : \boldsymbol{\theta}'_{*2} = \mathbf{c} \quad vs \quad H_1 : \boldsymbol{\theta}'_{*2} \neq \mathbf{c}, \tag{2.10}$$

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

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

$$LM(S) = S_2(\tilde{\boldsymbol{\theta}}_n)'\hat{A}_n^{22}(\tilde{\boldsymbol{\theta}}_n)^{-1}\hat{C}_{n22}(\tilde{\boldsymbol{\theta}}_n)^{-1}\hat{A}_n^{22}(\tilde{\boldsymbol{\theta}}_n)^{-1}S_2(\tilde{\boldsymbol{\theta}}_n), \tag{2.11}$$

where $\hat{A}^{22}(\tilde{\boldsymbol{\theta}}_n)$ is computed as in (2.6) 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 $H_1 : \boldsymbol{\theta}_{*2} = \boldsymbol{c} + \frac{\xi}{\sqrt{n}}$, 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)

$$\lambda = \frac{1}{n}\boldsymbol{\xi}' A_n^{22'} (B_{n22} - A_{n21}A_{n11}^{-1}B_{n12} - B_{n21}A_{n11}^{-1}A_{n12} + A_{n21}A_{n11}^{-1}B_{n11}A_{n11}^{-1}A_{n12})^{-1} A_n^{22}\boldsymbol{\xi}, \tag{2.12}$$

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 (2.6). All matrices in formula (2.12) are evaluated at $\boldsymbol{\theta}_*$.

Estimation procedure for the non-centrality parameter

The estimation methods described in section 2.3.1 to compute the asymptotic power are used here to estimate the asymptotic power for the LM(S) test, with some differences.

In step 3 of the first method, 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).

In step 4 of the second method, the non-centrality parameter is computed according to formula (2.12), where the matrices $A(\boldsymbol{\theta}^*)$ and $B(\boldsymbol{\theta}^*)$ are consistently estimated by $\hat{A}$ and $\hat{B}$, evaluated at $\tilde{\boldsymbol{\theta}}_n$.

Moreover, for both methods, 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) test have the same non-centrality parameter and, consequently, the same asymptotic power.

2.3.3 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 (2.10) is

$$GS = S_2(\tilde{\boldsymbol{\theta}})' V_{S_2}^{-1}(\tilde{\boldsymbol{\theta}}) S_2(\tilde{\boldsymbol{\theta}}), \tag{2.13}$$

where $S_2(\tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}$ are defined similarly as section 2.3.2, 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 (2.13) and (2.11) 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 (2.13), where $V_{S_2}(\tilde{\boldsymbol{\theta}})$ is estimated with the delete-1 Jackknife method as:

$$\hat{V}_{s_2}(\tilde{\boldsymbol{\theta}}_n) = \frac{n}{n-1} \sum_{i=1}^{n} (\tilde{S}_{2(i)} - \tilde{S}_2)(\tilde{S}_{2(i)} - \tilde{S}_2)'. \tag{2.14}$$

$\tilde{S}_{2(i)}$ 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.

2.4 Simulation study 1: The asymptotic power

The aim of this section is to compare the different procedures to estimate the asymptotic power of the LM tests, described in sections 2.3.1 and 2.3.2, by means of a small simulation study. We compare these procedures with the empirical power to study uniform DIF, considering only the LM(H) and LM(S) tests and the cases of correct model specification and misspecification of the latent variable distribution. The method that better estimates the asymptotic power will be applied in section 2.5 in other conditions of the study.

Both under correct and model misspecification, we consider a binary group variable x because we study measurement non-invariance only in two subgroups of population. Given n individuals and p items, under correct model specification, data are generated from the following model, where uniform DIF is introduced on the intercept of the last item p through the parameter γ_1 and the group variable x:

$$\begin{aligned} &logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i \qquad i = 1, ..., n \qquad j = 1, ..., p-1 \\ &logit(\pi_{ip}) = \alpha_{0p} + \alpha_{1p} z_i + \gamma_1 x_i \\ &z \sim N(0,1) \end{aligned} \tag{2.15}$$

Under misspecification of the latent variable distribution data are generated from the following model, where uniform DIF is specified as before:

$$\begin{aligned} &logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i \qquad i = 1, ..., n \qquad j = 1, ..., p-1 \\ &logit(\pi_{ip}) = \alpha_{0p} + \alpha_{1p} z_i + \gamma_1 x_i \\ &z \sim SN(\kappa) \end{aligned} \tag{2.16}$$

In this case, the latent variable z is generated from a skew-normal (SN) with skewness parameter κ, with the following probability density function (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 takes values from $-\infty$ to $+\infty$: when it is equal to 0, the skew-normal reduces to a standard normal distribution. In the simulations, we consider two values of κ, 3 and 5. When $\kappa = 3$ the mean and the variance of the latent variable are 0.76 and 0.43, respectively, and when $\kappa = 5$, the mean and the variance of the latent variable are 0.78 and 0.39, respectively. In both models (2.15) and (2.16) we consider two possible effect sizes, equal to 0.2 and 0.5, for the parameter γ_1. Moreover, in both cases, the values xs are generated from a Bernoulli distribution with success probability 0.7, the intercepts from a normal distribution with 0 mean and Standard Deviation (SD) 0.1 and the slopes from a normal distribution with 0 mean and SD 0.5.

The following set of hypotheses is being tested:

$$H_0 : \gamma_1 = 0 \quad vs \quad H_1 : \gamma_1 \neq 0 \tag{2.17}$$

Always the last item is tested for uniform DIF. Model (2.15) is fitted to the data with γ_1 fixed to 0. When data are generated from model (2.16) we are working under model misspecification. The following simulation conditions are considered: number of items $(p = 10)$ × sample size $(n = 200, 500, 1000, 5000, 10000)$ × Test statistic $(LM(H), LM(S))$. Due to the time complexity, the empirical power is computed only for $n = 200, 500, 1000$. 200 replications are considered for each condition of the study. The empirical power $\hat{p}$ is 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 the 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 95-th percentile of the χ^2_{df} distribution, with df equal to the number of constrained parameter under H_0. If non-valid statistics occur, they are excluded from the analysis. The asymptotic power is computed through methods 1 and 2 described in sections 2.3.1 and 2.3.2. The nominal level α is equal to 0.05 in all simulations. ML estimates of the parameters are obtained with direct maximization of the likelihood function using 21 Gauss-Hermite quadrature points. Numerical derivatives are used to compute the Hessian and cross-product matrices. Table 2.1 shows the results for the LM(H) and LM(S) tests computed under correct model specification when γ_1 is equal to 0.2 and 0.5 in the data generating model, $p = 10$, $n = 200, 500, 1000, 5000, 10000$, for the system of hypothesis (2.17).

TABLE 2.1: Asymptotic and empirical power of the LM(H) and LM(S) tests under correct model specification, $\gamma_1 = 0.2, 0.5$, $p = 10$, $n = 200, 500, 1000, 5000, 10000$.

			Method 1		Method 2		Empirical	
p	γ_1	n	LM(H)	LM(S)	LM(H)	LM(S)	LM(H)	LM(S)
10	0.2	200	0.086	0.085	0.080	0.079	0.08	0.06
		500	0.144	0.140	0.126	0.122	0.185	0.17
		1000	0.241	0.234	0.204	0.198	0.26	0.25
		5000	0.802	0.785	0.714	0.696	-	-
		10000	0.978	0.973	0.947	0.938	-	-
10	0.5	200	0.240	0.222	0.229	0.211	0.285	0.235
		500	0.508	0.468	0.484	0.445	0.54	0.5
		1000	0.799	0.758	0.775	0.732	0.8	0.78
		5000	1	1	1	1	-	-
		10000	1	1	1	1	-	-

We can notice that, in general, the differences between the asymptotic and empirical power are small and method 1 is slightly closer to the empirical power than method 2. For what concerns the power to detect measurement non-invariance, the LM(H) test has a slightly higher power compared to the LM(S) tests under all conditions, with the exception of the case $\gamma_1 = 0.5$ and for large sample sizes ($n = 5000, 10000$), where the two tests reach the same power, as expected from the theory. Table 2.2 shows the results for the LM(H) and LM(S) tests computed under misspecification of the latent variable distribution when γ_1 is equal to 0.2 and 0.5 in the data generating model, $p = 10$, $n = 200, 500, 1000, 5000, 10000$, for the system of hypothesis (2.17).

TABLE 2.2: Asymptotic and empirical power of the LM(H) and LM(S) tests under incorrect distribution of the latent variable, $\gamma_1 = 0.2, 0.5$, $p = 10$, $n = 200, 500, 1000, 5000, 10000$.

				Method 1		Method 2		Empirical	
p	ES	α	n	LM(H)	LM(S)	LM(H)	LM(S)	LM(H)	LM(S)
10	0.2	3	200	0.066	0.065	0.071	0.070	0.085	0.04
			500	0.091	0.089	0.104	0.101	0.11	0.075
			1000	0.133	0.129	0.159	0.154	0.185	0.14
			5000	0.464	0.447	0.569	0.550	-	-
			10000	0.753	0.734	0.854	0.839	-	-
		5	200	0.069	0.068	0.071	0.070	0.07	0.055
			500	0.010	0.097	0.102	0.010	0.135	0.085
			1000	0.151	0.146	0.157	0.151	0.145	0.135
			5000	0.538	0.517	0.561	0.540	-	-
			10000	0.828	0.809	0.848	0.829	-	-
10	0.5	3	200	0.158	0.145	0.170	0.155	0.202	0.13
			500	0.325	0.292	0.353	0.317	0.41	0.34
			1000	0.567	0.514	0.609	0.555	0.625	0.585
			5000	0.997	0.994	0.998	0.997	-	-
			10000	1	1	1	1	-	-
		5	200	0.163	0.148	0.168	0.153	0.21	0.15
			500	0.337	0.301	0.347	0.310	0.425	0.345
			1000	0.585	0.529	0.601	0.544	0.61	0.57
			5000	0.998	0.995	0.999	0.996	-	-
			10000	1	1	1	1	-	-

Also in this case the differences between the asymptotic and empirical power are small. For what concerns the power to detect measurement non-invariance under model misspecification, despite the fact that the LM(S) test is derived under model misspecification, the LM(H) test has the highest power under all conditions. The two tests reach the same power only when $\gamma_1 = 0.5$ and $n = 10000$. In both Tables and for both tests, the power increases with the sample size and the effect size of the parameter γ_1 and decreases when the model is misspecified.

Overall, the simulation study highlighted that the asymptotic power, computed through the two different approximation methods for the non-centrality parameter, is very close to the empirical power, also under model misspecification. Although the two procedures to compute the asymptotic power give similar results, we prefer the first method because it only requires the values of the LM test statistics to compute the non-centrality parameter. The second method requires additional quantities that may not be available from standard software. For these reasons, we consider the first method to compute the asymptotic power in section 2.5.

2.5 Simulation study 2: The LM tests

In this section we extensively study the performance of the LM(H), LM(CP), LM(S) 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 Appendix A.3. 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 sections 2.5.1 and 2.5.2, uniform and non-uniform DIF are studied in the simulation as well as single and multiple parameter hypotheses. We consider the following simulation conditions: number of items $(p = 10, 20)$ × 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 computed with the first procedure is reported 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, $R = 500$ 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 Appendix A, do we consider $R = 200$.

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 et al. (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 $\hat{p} \pm 1.96\sqrt{\frac{0.05(1-0.05)}{N_v}}$.

2.5.1 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:

$$\begin{aligned}
&logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i, \qquad i = 1, ..., n \qquad j = 1, ..., d,\ 1 \leq d \leq p \\
&logit(\pi_{iJ}) = \alpha_{0J} + \alpha_{1J} z_i + u_i, \qquad J = d+1, ..., p \qquad u \sim N(0, \sigma_u^2) \\
&z_i = \beta x_i + \epsilon_i \qquad \epsilon \sim N(0,1)
\end{aligned} \tag{2.18}$$

Both for $p = 10$ and for $p = 20$, the intercept parameters are generated from a log-normal distribution with mean 0 and standard deviation (SD) 0.1, the slope parameters are generated from a 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:

$$\begin{aligned}
&logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i, \qquad i = 1, ..., n \qquad j = 1, 2, ..., k \qquad 1 \leq k \leq p \\
&logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i + \gamma_{1j} x_i + \gamma_{2j} x_i z_i, \qquad j = k+1, ..., p \\
&z_i = \beta x_i + \epsilon_i, \qquad \epsilon \sim N(0,1),
\end{aligned} \tag{2.19}$$

where items $(k+1, ..., p)$ are tested for measurement invariance. In the case of uniform DIF, equation (2.19) 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 (2.19) 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 $H_0 : \gamma'_{1*} = \mathbf{0} \qquad vs \qquad H_1 : \gamma'_{1*} \neq \mathbf{0},$

where γ'_{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 $H_0 : \gamma'_{1*} = \mathbf{c} \qquad vs \qquad H_1 : \gamma'_{1*} \neq \mathbf{c}$, 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 2.3 presents the false positive rates for the LM(H), LM(CP), and LM(S) tests under local dependence for scenarios **A**, **B** and **C**.

TABLE 2.3: 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

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

TABLE 2.4: 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

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 2.5 presents the empirical power for the LM(H), LM(CP), and LM(S) tests under local dependence for scenarios **E** and **F**.

TABLE 2.5: 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

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 $LD = 50\%$.

2.5.2 Misspecification of the latent variable distribution

The data are generated from the following model:

$$\begin{aligned} &logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i \\ &z_i = \beta x_i + \epsilon_i, \qquad i = 1, ..., n \qquad j = 1, 2, ..., p \end{aligned} \tag{2.20}$$

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 mixture of normals considered does not represent a case of extreme outliers in the latent variable distributions.

Intercepts (α_{0j}), factor coefficients (α_{1j}), regression coefficient (β), and group variable x are generated as in section 2.5.1. Similarly here, we consider the model in equation (2.19) as the unconstrained model. The simulation scenarios of section 2.5.1 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 2.6 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**.

TABLE 2.6: 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

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 2.7 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**.

TABLE 2.7: 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

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

TABLE 2.8: 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

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.

2.6 Simulation study 3: The GS(J) test

In this section we study the performance of 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 (2.14) 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 $(p = 10)$ × sample size $(n = 200, 500, 1000)$ × 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 (2.18), and iii) under misspecification of the latent variable distribution from the model given in equation (2.20). 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, (2.18), (2.20)). 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 section 2.5.1. When data are generated from (2.18), we consider $\sigma_u^2 = 1$ and $LD = 20\%$. For data generated from (2.20), we assume $\epsilon \sim SN(3)$. We consider the model in equation (2.19), without the structural model, as the unconstrained model. Under scenario **A** of section 2.5.1, 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 2.9 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**.

TABLE 2.9: Type I error/ false positive rates 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
(2.18)	**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
(2.20)	**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

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 2.10 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**.

TABLE 2.10: 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
(2.18)	**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
(2.20)	**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

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.

2.7 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 et al. (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

$H_0 : \gamma_{2j*} = 0 \;\; vs \;\; H_1 : \gamma_{2j*} \neq 0$, 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 section 2.5.2, 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 2.11 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.

TABLE 2.11: 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

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.

2.8 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 various simulation studies 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. It should be mentioned that considering a mixture of normals that represents the case of extreme outliers, as the one in Ma and Genton (2010), could give similar results in terms of tests performance as the ones obtained when the true latent variable is highly skewed.

The simulation studies highlight that there are small numerical differences between the asymptotic power, computed through the two different approximation methods 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. Among the two procedures to compute the asymptotic power we prefer the first method because it only requires the value of the test statistic to compute the non-centrality parameter. Although not shown here, the asymptotic power can be used also to find sample sizes necessary to reach a certain power (Boos and Stefanski, 2013, Gudicha et al., 2017).

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.

Chapter 3

The Generalized Hausman test to detect non-normality of the latent variable distribution

3.1 Introduction

One of the typical assumptions of latent variable models is the normal distribution of the latent variable(s). As shown in Ma and Genton (2010), this assumption is not always appropriate and misspecifying the form of the latent variable by assuming normality can result in large biases in parameter estimates. Moreover, as shown in Chapter 2, an incorrect distribution of the latent variable can lead to incorrect inference when performing hypothesis testing. Assuming a different form for the latent variable is not new in the literature of the generalized linear latent variable models (GLLVM) and IRT models. For example Montanari and Viroli (2010) introduce a skew-normal latent variable in the factor model, while Cagnone and Viroli (2012) present a latent trait model where the factors are distributed as a finite mixture of multivariate gaussians. Ma and Genton (2010) propose a semiparametric method for GLLVM, consistent for various types of manifest variables under different distributions of the latent variables. Irincheeva et al. (2012) consider the seminonparametric (SNP) approach, introduced by Gallant and Nychka (1987), within the GLLVM framework. This approach allows for more flexible smooth densities of the latent variables. In IRT different methods have been proposed to deal with non-normal latent variable(s). For example, for binary responses, Knott and Tzamourani (2007) use the empirical histogram method combined with the bootstrap, estimating the prior distribution of the latent variable from the data instead of assuming standard normality. Woods (2006) proposes the so-called Ramsey curve IRT model, where the latent variables are splines based densities, that are linear combination of polynomial functions joint together at knots. This method implies a modification of the standard E-M algorithm. The SNP method has been used in unidimensional IRT model by Woods and Lin (2009) and in multidimensional IRT model by Monroe (2014).

Different methods have been proposed to detect non-normality of the latent variables. For example, with continuous manifest variables, Ma and Genton (2010) perform the Kolmogorov–Smirnov test on the normality of the latent variable by inspecting the continuous responses. Commonly information criteria are used to choose between a model where the latent variables are normally distributed and a model where the latent variables have a more complex shape (Woods and Lin, 2009, Irincheeva et al., 2012, Monroe, 2014).

Hausman (1978) proposes a specification test to detect failure of the orthogonality assumption of the regression model. The Hausman test can be applied also in other contexts, to detect different types of model misspecification, and it can be used as an alternative to the classic information criteria (Bartolucci et al., 2017). The idea of the test is simple. It compares two different estimators that are consistent when the model is correctly specified and one is also efficient. However, in the presence of model misspecification, only the inefficient estimator is consistent. The efficiency assumption simplifies the computation of the covariance matrix of the difference of the two estimators. However, this covariance matrix can fail to be positive definite in presence of model misspecification. Moreover, sometimes none of the two estimators considered are fully efficient.

For these reasons, we consider a generalized version of the Hausman test, proposed by White (1982), and we refer to this test as Generalized Hausman (GH) test. To build the GH test none of the two estimators need to be fully efficient and the covariance matrix of the difference of the two estimators is robust and always positive definite. This test compares a classic maximum-likelihood (ML) estimator, that is inconsistent under model misspecification, with a Quasi-ML estimator, consistent under correct and model misspecification for a particular set of parameters.

In the IRT context, as far as we know, the classic Hausman test has been used only by Ranger and Much (2020) to detect misspecification of the item characteristic functions and local dependencies among items. They implement two versions of the test, one for the local and one for the global fit of the model. For the global fit, they compare the classic Hausman test with the M_2 test. They highlight that the M_2 test has always good performance in terms of Type I error rates, while the Hausman test only for large sample sizes. The Hausman test has lower power to detect local dependence than the M_2 test. The latter, however, has very low or no power when the misspecification is in the form of an upper boundary item characteristic function. The M_2 and Hausman tests have the same power when the misspecification is caused by a non-monotone item characteristic function.

In generalized linear mixed models (GLMM) for clustered data, a robust version of the Hausman test, similar to the one by White (1982), has been proposed by Bartolucci et al. (2017) when a discrete distribution for the random effects is assumed. The test can be also used to detect the possible correlation between random effects and cluster-specific covariates. With respect to the information criteria, they found that the robust Hausman test prefers more parsimonious models, under a true continuous distribution of the random effects and a correct specification of the dependence between the random effects and the covariates, and it can detect the presence of endogeneity, ignored by information criteria.

The objective of this Chapter is to extend the GH test to detect non-normality of the latent variable distribution in unidimensional IRT models for binary data. In order to apply the test, we consider the estimators of two different models. The first one is the ML estimator that maximizes the likelihood of a classic unidimensional IRT model for binary data based on the normality assumption of the latent variable. This estimator is consistent under the null hypothesis of normality of the latent variable but produces biased parameter estimates when the latent variable is not normally distributed (Ma and Genton, 2010). The second one is the quasi-ML estimator of the unidimensional SNP-IRT model for binary data (Woods and Lin, 2009, Irincheeva, 2011), consistent under the normality and under different distribution assumptions of the latent variable (Gallant and Tauchen, 1989, Irincheeva et al., 2012). Once the non-normality of the latent variable distribution is identified, hypothesis testing should be carried out with the correct estimation method.

We carry out a simulation study to evaluate the performance of the GH test to detect non-normality of the latent variable in unidimensional IRT model for binary data, in terms of Type I error rates and empirical power, varying the distribution assumptions of the latent variable. Similarly to Ranger and Much (2020), the performance of the GH test is compared with the M_2 test. Since the SNP-IRT model and the classic IRT model for binary data are nested, also the Likelihood-Ratio (LR) test is considered in the simulations and, in addition, some information criteria are computed. Two applications to real data are also presented.

The Chapter is organized as follows. First, we present the unidimensional SNP-IRT model for binary data. Second, we review the information criteria and the M_2 test. Third, we introduce the GH test to detect non-normality of the latent variable distribution. Next, we present a Monte Carlo simulation study and the results from two real data analysis. Finally, we present some concluding remarks.

3.2 The SNP-IRT model

3.2.1 The model

As in Chapter 2, let us denote by $y_1, ..., y_p$ a set of observed binary variables/items, by z the latent variable. The response probability for the i-th individual to the j-th item is modelled using a logistic model (measurement model) where the latent variable has a SNP parametrization:

$$\begin{aligned} P(y_{ij} = 1|z_i) &= \pi_{ij}(z_i) = \frac{\exp\left(\alpha_{0j} + \alpha_{1j} z_i\right)}{1 + \exp\left(\alpha_{0j} + \alpha_{1j} z_i\right)} \\ h(z_i) &= P_L^2(z_i)\phi(z_i) \qquad P_L(z_i) = \sum_{0 \le l \le L} a_i z_i^l. \end{aligned} \tag{3.1}$$

$a_0 ... a_L$ are the real coefficients of the polynomial $P_L(z_i)$, L is the polynomial degree and $\phi(z_i)$ is the density of a standard normal. The case $L = 0$ corresponds to $Z \sim N(0, 1)$.

In order for $h(z)$ to be a density, the coefficients $a_0, ..., a_q$ of $P_L(z)$ should be chosen such that $\int h(z)dz = 1$. For this purpose, Gallant and Tauchen (1989) use a proportionality constant $1/\int P_L(z)^2\phi(z)dz$ and fix the constant term of the polynomial equal to 1. Alternatively, Irincheeva et al. (2012) and Woods and Lin (2009) use the parametrization proposed by Zhang and Davidian (2001), that imposes

$$1 = \int_R P_L^2(z)\phi(z)dz = E\{P_L^2(w)\} = a'E(\tilde{w}\tilde{w}')a = a'Aa \tag{3.2}$$

with $w \sim N(0, 1)$, $P_L(w) = a'\tilde{w}$, $\tilde{w} = (1, w, w^2, ..., w^L)$. The matrix A is positive definite by definition and $A = B'B$, where B is a positive definite matrix.

If $c = Ba$, equation (3.2) becomes $c'c = 1$ and $c = (c_1, ..., c_{L+1})'$. The elements of c can be represented using a polar coordinate transformation as $c_1 = \sin\varphi_1, c_2 = \cos\varphi_1 \sin\varphi_2, ..., c_L = \cos\varphi_1 ... \cos\varphi_{L-1} \sin\varphi_L, c_{L+1} = \cos\varphi_1 \cos\varphi_2 ... \cos\varphi_{L-1} \cos\varphi_L$, with $-\pi/2 < \varphi_t \leq \pi/2, t = 1, ..., L$. The density of the latent variable in (3.1) can be expressed as

$$h(z|\boldsymbol{\varphi}, L) = (a'\tilde{\mathbf{z}})^2\phi(z), \tag{3.3}$$

where a can be obtained from c as $a = B^{-1}c, \tilde{\mathbf{z}} = (1, z, z^2, ..., z^L)'$ and $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_L)'$.

When $L = 2$, $P_L(z) = a_0 + a_1 z + a_2 z^2$, $a_0 = \sin\varphi_1 - \frac{1}{\sqrt{2}}\cos\varphi_1\cos\varphi_2$, $a_1 = \cos\varphi_1 \sin\varphi_2$ and $a_2 = \frac{1}{\sqrt{2}}\cos\varphi_1\cos\varphi_2$. If we set $\varphi_2 = \frac{\pi}{2}$, we get the coefficients for $L = 1$, $P_L(z) = a_0 + a_1 z$, $a_0 = \sin\varphi_1$, $a_1 = \cos\varphi_1$. When $L = 0$ the model in (3.1) reduces to the classic IRT model for binary data with a standard normal distributed latent variable. In the following sections we indicate with SNP_2 the model for $L = 2$, with SNP_1 for $L = 1$ and with SNP_0 for $L = 0$.

For a random sample of size n the log-likelihood is:

$$\begin{aligned} l(\mathbf{y}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln f(\mathbf{y}_i, \boldsymbol{\theta}) = \\ &= \sum_{i=1}^{n} \ln \int \prod_{j=1}^{p} \pi_{ij}(z_i)^{y_{ij}}(1 - \pi_{ij}(z_i))^{1-y_{ij}} P_L^2(z_i) \exp\left(-\frac{1}{2}z_i'z_i\right) dz_i, \end{aligned} \tag{3.4}$$

where $\boldsymbol{\theta}$ is the vector of the unknown parameters that includes item intercepts, slopes and $\boldsymbol{\varphi}$ parameter. The integral in the log-likelihood $l(\mathbf{y}, \boldsymbol{\theta})$ is approximated with the Gauss-Hermite quadrature, as in Woods and Lin (2009). The degree of the polynomial L is fixed and is not estimated by maximum likelihood. The log-likelihood function is maximized with respect to the unknown vector of parameter $\boldsymbol{\theta}$ as follows

$$\boldsymbol{\theta}^* = (\boldsymbol{\alpha}_0^*, \boldsymbol{\alpha}_1^*, \boldsymbol{\varphi}^*) = argmax_{\boldsymbol{\theta}} l(\mathbf{y}, \boldsymbol{\theta}) \tag{3.5}$$

The final estimator is obtained rescaling $\boldsymbol{\theta}^*$ for identifiability reasons, as described in the next section.

3.2.2 The identifiability problem

The standard argument for obtaining the final form of the estimators $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\alpha}}_1$ relies on the concept of indeterminacy of the scale of the latent variable (Lord, 1980, van der Linden and Barrett, 2016). Indeed, in latent variable models, it is possible to change the scale of the latent variable z changing the parametrization accordingly and obtain the same joint density function of the observed variables. For this reason, we cannot expect that the latent variable in the SNP-IRT model has the same mean and variance of the true latent variable, but we expect that it has the same shape. The latent variable z has density $h(z|\boldsymbol{\varphi}^*, L)$ with estimated mean $\tilde{E}(Z)$ and variance $\tilde{V}(Z)$ that can be different from the true mean and variance of the latent variable. In order to compare the classic IRT model with the SNP-IRT model, we fix the mean and the variance of the latent variable of the SNP-IRT model to 0 and 1, respectively. After the optimization process, we can express

$$E(logit(\pi_j(z)) = \alpha_{0j}^* + \alpha_{1j}^* z \qquad j = 1, ..., p \tag{3.6}$$

and

$$z = \sqrt{\tilde{V}(Z)} z_1 + \tilde{E}(Z) \tag{3.7}$$

where $\tilde{E}(Z)$ and $\tilde{V}(Z)$ are found given $\boldsymbol{\varphi}^*$ and the SNP density of z. z_1 has the same distribution of z, with mean 0 and variance 1.

If we substitute (3.7) in (3.6) we get

$$E(logit(\pi_j(z)) = \alpha^*_{0j} + \alpha^*_{1j}\sqrt{\tilde{V}(Z)}z_1 + \alpha^*_{1j}\tilde{E}(Z) \qquad j = 1, ..., p \tag{3.8}$$

From equation (3.8) we get the form of the final estimators (Irincheeva et al., 2012):

$$\hat{\alpha}_{0j} = \alpha^*_{0j} + \alpha^*_{1j}\tilde{E}(Z) \qquad j = 1, ..., p \tag{3.9}$$

$$\hat{\alpha}_{1j} = \alpha^*_{1j}\sqrt{\tilde{V}(Z)} \qquad j = 1, ..., p \tag{3.10}$$

$\tilde{E}(Z)$ and $\tilde{V}(Z)$ can be computed analytically, given the values of $\boldsymbol{\varphi}^*$, as shown in the Appendix B. The standard errors of $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\alpha}}_1$ are based on the observed Hessian and cross-product matrix and they can be obtained with the Delta method (Cramér, 1946). In the simulations, if the mean of the true latent variable is different from 0 and the variance from 1, to compute the parameter bias, the true $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ parameters are rescaled accordingly to formula (3.9) and (3.10), replacing α^*_{0j} with the true intercept, α^*_{1j} with the true slope and $\tilde{E}(Z)$ and $\tilde{V}(Z)$ with the mean and the variance of the true latent variable. In this way, the parameter bias is due only to the misspecification of the shape of the latent variable, and not to the misspecification of the moments. A similar procedure to compute parameter bias has been adopted also by Irincheeva (2011).

If more than one latent variable is considered in the SNP-IRT model, other parameter constraints are needed for model identifiability (see Irincheeva, 2011).

3.2.3 Model selection criteria

The Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the Hannan–Quinn criterion (HQ) can be used to choose the degree of the polynomial L of the SNP-IRT model (Davidian and Gallant, 1993, Woods and Lin, 2009, Irincheeva et al., 2012, Monroe, 2014).

The AIC is (Akaike, 1974):

$$AIC = -2l(\mathbf{y}, \boldsymbol{\theta}^*) + 2k, \tag{3.11}$$

where $l(\mathbf{y}, \boldsymbol{\theta}^*)$ is the log-likelihood defined in (3.4) and evaluated at $\boldsymbol{\theta}^*$ and k is the number of parameter of the model.

The BIC is (Schwarz, 1978):

$$BIC = -2l(\mathbf{y}, \boldsymbol{\theta}^*) + k \ln n, \tag{3.12}$$

where n is the sample size.

The HQ is (Hannan and Quinn, 1979):

$$HQ = -2l(\mathbf{y}, \boldsymbol{\theta}^*) + 2k \ln \ln n \tag{3.13}$$

Usually $L = 1$ or $L = 2$ are enough to detect a departure from normality and to approximate different shapes of latent variable distributions. Selecting higher order of the polynomial could result in overfitting (Irincheeva, 2011).

3.3 The M_2 test

In the case of normality of the latent variable, the overall goodness-of-fit of the SNP_0 model is usually assessed through the classic Pearson's chi-square test and Likelihood-Ratio test (Bartholomew et al., 2011). However, these test are affected by the problem of sparse data. Indeed, the number of empty cells in a frequency table increases with the number of binary items. In this case, the distribution of these classical statistics is not well approximated by the chi-square distribution. To overcome this problem, Maydeu-Olivares and Joe (2005) propose a family of test statistics M_r, based on the residuals up to order r. The most popular statistic is M_2, that uses the univariate and bivariate marginal information. As data sparseness increases, the empirical Type I error rates of the M_2 test remain accurate (Maydeu-Olivares and Joe, 2005, 2006). As before, let's consider p items and a sample size n. Under the null hypothesis we test that the SNP_0 model holds. The hypotheses H_0 and H_1 can be formalized as follows:

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}) \quad vs \quad H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta}), \tag{3.14}$$

where $\boldsymbol{\theta}$, as usual, includes the item intercepts and slopes and $\boldsymbol{\pi}(\boldsymbol{\theta})$ indicates the response patterns probabilities.

The statistic M_2 is (Maydeu-Olivares and Joe, 2005):

$$M_2 = n\hat{\mathbf{e}}_2'\hat{\boldsymbol{U}}_2\hat{\mathbf{e}}_2. \tag{3.15}$$

The vector $\hat{\mathbf{e}}$ includes the univariate and bivariate residuals and is computed as

$$\hat{\mathbf{e}}_2 = (\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}})), \tag{3.16}$$

where $\boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}})' = (\dot{\boldsymbol{\pi}}_1(\hat{\boldsymbol{\theta}})', \dot{\boldsymbol{\pi}}_2(\hat{\boldsymbol{\theta}})')$. $\dot{\boldsymbol{\pi}}_1(\hat{\boldsymbol{\theta}})$ includes the p estimated first-order marginal probabilities of correctly responding to each single item and $\dot{\boldsymbol{\pi}}_2(\hat{\boldsymbol{\theta}})$ includes the $\frac{p(p-1)}{2}$ estimated second-order marginal probabilities of correctly responding to each item pair. $\hat{\boldsymbol{\theta}}$ is the vector of ML parameter estimates under the SNP_0 model. It is possible to compute $\boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}})$ as:

$$\boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}) = T_2\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}), \tag{3.17}$$

where T_2 is a transformation matrix of 1s and 0s of dimension $\frac{p(p+1)}{2} \times 2^p$, where $\frac{p(p+1)}{2}$ is the number of univariate and bivariate residuals (for more details on this matrix see Maydeu-Olivares and Joe, 2005) and $\boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ is the matrix of the estimated probabilities of the observed response patterns. Similarly to $\boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}})$, $\boldsymbol{p}_2$ includes the observed cell proportions of 1 for each single item and each pair of items.

The matrix $\hat{\boldsymbol{U}}_2$ in (3.15) has dimension $\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}$ and it is computed as:

$$\hat{\boldsymbol{U}}_2 = \hat{\boldsymbol{\Xi}}_2^{-1} - \hat{\boldsymbol{\Xi}}_2^{-1}\hat{\boldsymbol{\Delta}}_2(\hat{\boldsymbol{\Delta}}_2'\hat{\boldsymbol{\Xi}}_2^{-1}\hat{\boldsymbol{\Delta}}_2)^{-1}\hat{\boldsymbol{\Delta}}_2'\hat{\boldsymbol{\Xi}}_2^{-1}, \tag{3.18}$$

where $\hat{\boldsymbol{\Xi}}_2 = T_2\hat{\boldsymbol{\Xi}}T_2'$ and $\hat{\boldsymbol{\Xi}} = diag(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))$. The matrix $\hat{\boldsymbol{\Delta}}_2$ is computed as $\hat{\boldsymbol{\Delta}}_2 = T_2\hat{\boldsymbol{\Delta}}$, where $\hat{\boldsymbol{\Delta}}$ is the Jacobian matrix of derivatives of the cell probabilities with respect to the items intercept and slope parameter.

Under H_0, the statistic M_2 is asymptotically distributed as a χ^2_m, with degrees of freedom $m = \frac{p(p+1)}{2} - 2p$, that is the number of univariate and bivariate residuals minus the number of estimated parameters of the SNP_0 model.

In the simulations, we evaluate the performance of the M_2 test also under non-normality of the latent variable, expecting the test to reject the null hypothesis that the SNP_0 model holds.

3.4 The Generalized Hausman test

In this section we present the GH test, derived by White (1982), applied to detect non-normality of the latent variable using the SNP-IRT model.

Let's denote by $\boldsymbol{\eta}$ the sub-vector of $\boldsymbol{\theta}' = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', \boldsymbol{\varphi}')$ that includes the item intercepts $\boldsymbol{\alpha}_0$ and slopes $\boldsymbol{\alpha}_1$. $\boldsymbol{\eta}$ has dimension $2p \times 1$, where p is the number of items.

Let's consider the ML-estimator $\hat{\boldsymbol{\theta}}_{SNP_0} = \hat{\boldsymbol{\eta}}_{SNP_0}$, that maximizes the likelihood function of a classic IRT model where the latent variable is normally distributed, that is the SNP_0 model. Under normality of the latent variable $\hat{\boldsymbol{\eta}}_{SNP_0} \xrightarrow{p} \boldsymbol{\eta}_0$, where $\boldsymbol{\eta}_0$ includes the true item intercepts and slopes. Under non-normality of the latent variable, $\hat{\boldsymbol{\eta}}_{SNP_0} \xrightarrow{p} \boldsymbol{\eta}_*$, where $\boldsymbol{\eta}_*$ is the vector that minimizes the Kullback-Leibler information criterion.

Let's also consider a Quasi-ML estimator $\hat{\boldsymbol{\theta}}'_{SNP_L} = (\hat{\boldsymbol{\eta}}'{}_{SNP_L}, \boldsymbol{\varphi}^{*'})$, that maximizes the likelihood function of a SNP-IRT model with $L > 0$, where the sub-vector of parameter $\boldsymbol{\varphi}^*$ has dimension $L \times 1$ and so $\hat{\boldsymbol{\theta}}_{SNP_L}$ has dimension $(2p + L) \times 1$. Under normality and different distributional assumptions of the latent variables and if the regularity conditions A2-A6 of White (1982) are satisfied, $\hat{\boldsymbol{\theta}}_{SNP_L} \xrightarrow{p} \boldsymbol{\theta}_{0*}$, where $\boldsymbol{\theta}'_{0*} = (\boldsymbol{\eta}_0', \boldsymbol{\varphi}_*{}')$ and $\boldsymbol{\varphi}_*$ is the value of $\boldsymbol{\varphi}$ that minimizes the Kullback-Leibler information criterion (White, 1982, Gallant and Tauchen, 1989, Irincheeva et al., 2012). This assumption ensures that $\hat{\boldsymbol{\eta}}_{SNP_L}$ is a consistent estimator for the true set of parameter values $\boldsymbol{\eta}_0$, that include the true item intercepts and slopes, even under non-normality of the latent variable.

Following White (1982), under normality of the latent variable

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{SNP_L} - \hat{\boldsymbol{\eta}}_{SNP_0}) \xrightarrow{d} N(0, S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})). \tag{3.19}$$

An estimator of $S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})$ is:

$$\hat{S}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L}) = \hat{A}^{\boldsymbol{\eta\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1}\hat{B}(\hat{\boldsymbol{\theta}}_{SNP_L})\hat{A}^{\boldsymbol{\eta\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1'} + \hat{A}(\hat{\boldsymbol{\eta}}_{SNP_0})^{-1}\hat{B}(\hat{\boldsymbol{\theta}}_{SNP_0})\hat{A}(\hat{\boldsymbol{\eta}}_{SNP_0})^{-1'} - \\ - \hat{A}^{\boldsymbol{\eta\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1}\hat{R}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})'\hat{A}(\hat{\boldsymbol{\eta}}_{SNP_0})^{-1'} - \hat{A}(\hat{\boldsymbol{\eta}}_{SNP_0})^{-1}\hat{R}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})\hat{A}^{\boldsymbol{\eta\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1'}, \tag{3.20}$$

where the matrices $\hat{A}(\hat{\boldsymbol{\eta}}_{SNP_0})$ and $\hat{B}(\hat{\boldsymbol{\eta}}_{SNP_0})$ are the observed Hessian and cross-product matrix of dimension $2p \times 2p$ for the model SNP_0 evaluated at $\hat{\boldsymbol{\eta}}_{SNP_0}$ and $\hat{A}(\hat{\boldsymbol{\theta}}_{SNP_L})$ and $\hat{B}(\hat{\boldsymbol{\theta}}_{SNP_L})$ are the observed Hessian and cross-product matrix of dimension $(2p + L) \times (2p + L)$ for the model SNP_L evaluated at $\hat{\boldsymbol{\theta}}_{SNP_L}$. The matrix $\hat{A}^{\boldsymbol{\eta\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1}$ is obtained deleting the last L row from the matrix $\hat{A}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1}$ and has dimension $2p \times (2p + L)$. The matrix $\hat{R}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ has dimension $2p \times (2p + L)$ and can be computed as:

$$\hat{R}(\boldsymbol{\eta}_{SNP_0}, \boldsymbol{\theta}_{SNP_L}) = \sum_{i=1}^{n} \frac{\partial l_{SNP_0}(\boldsymbol{y}_i, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \frac{\partial l_{SNP_L}(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \tag{3.21}$$

where $l_{SNP_0}(\boldsymbol{y}_i, \boldsymbol{\eta})$ is the log-likelihood for the individual i under the model SNP_0 and $l_{SNP_L}(\boldsymbol{y}_i, \boldsymbol{\theta})$ is the log-likelihood for the individual i under the model SNP_L. The matrix in (3.21) is evaluated at $(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$.

Given the theoretical result in (3.19), the GH test to detect non-normality of the latent variable is:

$$GH = (\hat{\boldsymbol{\eta}}_{SNP_L} - \hat{\boldsymbol{\eta}}_{SNP_0})'\hat{S}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})^{-1}(\hat{\boldsymbol{\eta}}_{SNP_L} - \hat{\boldsymbol{\eta}}_{SNP_0}) \tag{3.22}$$

Under normality of the latent variable, the GH test is asymptotically distributed as a χ^2_{2p}, where $2p$ are the degrees of freedom.

3.4.1 Approximation to the distribution of a quadratic form

The matrix $\hat{S}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ in (3.22) can be often close to singularity and its inversion can be numerically unstable.

Given the theoretical result in (3.19) and the quadratic form $(\hat{\boldsymbol{\eta}}_{SNP_L} - \hat{\boldsymbol{\eta}}_{SNP_0})'(\hat{\boldsymbol{\eta}}_{SNP_L} - \hat{\boldsymbol{\eta}}_{SNP_0})$, we consider the following test statistic (Yuan and Bentler, 2010, Ranger and Much, 2020):

$$GH_T = (\hat{\boldsymbol{\eta}}_{SNP_L} - \hat{\boldsymbol{\eta}}_{SNP_0})'(\hat{\boldsymbol{\eta}}_{SNP_L} - \hat{\boldsymbol{\eta}}_{SNP_0}) \tag{3.23}$$

Under normality of the latent variable

$$GH_T \sim \sum_{l=1}^{d} \lambda_i z_i^2, \qquad z_i \sim N(0,1). \tag{3.24}$$

It is possible to approximate the distribution in (3.24) with (Welch, 1938,Yuan and Bentler, 2004):

$$GH_T \sim a\chi_b^2. \tag{3.25}$$

The quantity a and b are defined as:

$$a = \frac{\sum_{l=1}^{d} \lambda_l^2}{\sum_{l=1}^{d} \lambda_l} \tag{3.26}$$

and

$$b = \frac{(\sum_{l=1}^{d} \lambda_l)^2}{\sum_{l=1}^{d} \lambda_l^2}, \tag{3.27}$$

where d is the rank of $S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})$ and $\lambda_1, ..., \lambda_d$ are its non-zero eigenvalues. Since $S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})$ can be consistently estimated by $\hat{S}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ defined in (3.20), a in (3.26) and b in (3.27) can be consistently estimated with

$$\hat{a} = \frac{\sum_{l=1}^{d} \hat{\lambda}_l^2}{\sum_{l=1}^{d} \hat{\lambda}_l} \tag{3.28}$$

and

$$\hat{b} = \frac{(\sum_{l=1}^{d} \hat{\lambda}_l)^2}{\sum_{l=1}^{d} \hat{\lambda}_l^2}, \tag{3.29}$$

where d is rank of $\hat{S}(\hat{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ and $\hat{\lambda}_1, ..., \hat{\lambda}_d$ are its non-zero eigenvalues.

3.5 Simulation study: The GH_T test

3.5.1 Design

In this section we study the performance of the GH_T test to assess non-normality of the latent variable distribution and we compare its performance with the M_2 and LR

tests. Moreover, for all simulation scenarios, the BIC, AIC and HQ criteria are computed. The estimation of the SNP-IRT model is computationally expensive. Moreover, as the degree of the polynomial L increases ($L > 1$), the SNP_L model becomes more sensitive to the choice of the initial values for all model parameters and the estimation results can be less reliable. Furthermore, we consider non-normal distribution for the latent variable in the data generating models, such as different mixtures of two normals, that can be well approximated with $L = 1$, as highlighted in the results on the parameter bias in Irincheeva et al. (2012). For these reasons, to implement the GH_T test, we consider the SNP_0 model and the SNP_1 model. We also consider skew-normal distributions in the data generating models, even if the SNP method approximates better the shape of a latent variable distributed as a mixture of normals (Monroe, 2014).

The optimization is achieved in R with direct maximization via the function “nlminb”, that uses the analytically computed gradient and Hessian matrix reported in the Appendix B. For the SNP_1 model, the initial values of the parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ used in the optimization process are the parameter estimates obtained with the SNP_0 model. In each data replication, for the φ_1 parameter, we sample 10 initial values from a sequence of values, equally spaced by 0.1 in the interval $[-\frac{\pi}{2}; \frac{\pi}{2}]$, i.e. the domain of φ_1. Among the estimated SNP_1 models in each data replication, the optimal one corresponds to the maximum value of the log-likelihood function. The Hessian and cross-product matrices involved in the GH_T test are computed analitically with the derivatives reported in the Appendix B and using the Delta method. The matrix in formula (3.21) is computed numerically with the “NumDeriv” R package. The M_2 test in each data replication is computed using the "M2" function of the R package "mirt".

We consider the following simulation conditions: number of items ($p = 4, 10, 20$) × sample size ($n = 200, 500, 1000$)× test statistics (GH_T, M_2, LR)× information criterion (AIC, BIC, HQ). In all the simulation scenarios, $R = 500$ replications are considered and two nominal levels of α are considered, that is $\alpha = 0.05, 0.01$. Non-valid statistics, for example negative statistics, are excluded from the analysis. The Type I error rates and power of the GH_T, M_2 and LR tests 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. c is the theoretical asymptotic critical value corresponding to the $(1-\alpha)$th percentile of the $\hat{a}\chi^2_{\hat{b}}$ distribution for the GH_T test, where $\hat{a}$ and $\hat{b}$ are computed as in (3.28) and (3.29) and of the χ^2_m for the M_2 test, where m is equal to $\frac{p(p+1)}{2} - 2p$. Since the SNP_1 model has one more parameter than the SNP_0 model, for the LR test c is the theoretical asymptotic critical value corresponding to the $(1-\alpha)$th percentile of the χ^2_1. The confidence interval (CI) of each rate $\hat{p}$ is computed as $\hat{p} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\alpha(1-\alpha)}{N_v}}$. The percentages of times the AIC, BIC and HQ criteria select the SNP_1 model instead of the SNP_0 model are computed as $\hat{P} = \sum_{l=1}^{N_v} \frac{I(IC_{SNP_1} < IC_{SNP_0})}{N_v} 100$, where IC indicates the AIC, BIC and HQ criterion.

To evaluate the performance of the GH_T, M_2 and LR tests, we consider three scenarios (SC), corresponding to three different distribution assumptions for the latent variable z in the data generating models. The general model is

$$\begin{aligned} &logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i \qquad i = 1, ..., n \qquad j = 1, 2, ..., p \\ &z \sim f(z) \end{aligned} \tag{3.30}$$

Item intercepts are randomly chosen from the interval [-0.8; 1.12] while the item slopes from the interval [0.5; 1.5].

To study the Type I error rates of the GH_T, M_2 and LR tests we consider the following scenario:

A1 $z \sim N(0,1)$

To study the power of the GH_T, M_2 and LR tests we consider the following two scenarios:

B1 $z \sim 0.1N(-2, 0.25) + 0.9N(2, 1)$,

where z has an overall mean equal to 1.6 and variance equal to 2.365.

C1 $z \sim 0.7N(-1.5, 0.6) + 0.3N(1.5, 0.5)$,

where z has an overall mean equal to -0.6 and variance equal to 2.217.

D1 $Z \sim SN(1.27)$,

where z has mean 0.6 and variance 3.864.

E1 $Z \sim SN(-1.58)$,

where z has mean -0.46 and variance 4.63.

The mixture of normals in the scenario **C1** wider departs from the normal distribution than the one in **B1**. The distribution reported in scenario **D1** is right-skewed, while the one in scenario **E1** is left-skewed. Moreover, the distribution in scenario **E1** largely departs from the normal one compared with the distribution in scenario **D1**.

3.5.2 Results

Table 3.1 reports the Type I error rates of the GH_T, M_2 and LR tests for scenario **A1**.

TABLE 3.1: Type I error rates of the GH_T, M_2 and LR tests for scenario A1, $p = 4, 10, 20$, $n = 200, 500, 1000$

		$\alpha = 0.05$			$\alpha = 0.01$		
p	n	GH_T	M_2	LR	GH_T	M_2	LR
4	200	**0.012**	0.056	0.05	0.008	0.012	0.004
	500	**0**	0.04	**0.092**	0	0.014	0.016
	1000	0.046	0.048	**0.103**	0.006	0.006	0.012
10	200	0.054	0.036	**0.161**	0.020	0.006	**0.03**
	500	0.043	0.066	0.068	0.020	0.012	0.018
	1000	0.034	0.046	**0.021**	0.02	0.01	0.004
20	200	0.048	0.06	**0.158**	0.02	0.008	**0.065**
	500	0.04	0.036	**0.078**	0.017	0.01	**0.023**
	1000	0.046	0.048	0.046	**0.036**	0.01	0.006

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

Overall, the GH_T test has good performance in terms of Type I error rates for the different levels of α considered. Indeed, for $\alpha = 0.05$, GH_T rejects less than expected only with 4 items, small and medium sample sizes. For $\alpha = 0.01$, it has a slightly inflated Type I error rate only with 20 items and large sample size. The M_2 has good performance in terms of Type I error rates for all values of α, number of items and sample sizes. Among the three tests considered, the LR test has the worst performance. For $\alpha = 0.05$, LR has inflated Type I error rates with 4 items and large sample sizes and with 10 and 20 items and small sample sizes. It also rejects less than expected in the case of 10 items and large sample size. For $\alpha = 0.01$, it has inflated Type I error rates with 10 items, small sample size and with 20 items, small and medium sample sizes. It is worth noting that, in some cases, the GH_T test produces non-valid results, ranging from 0.8% to 4.2% of the replications, where the highest percentage corresponds to 10 items and medium sample size. Also the LR test produces non-valid results, ranging from 0.8% to 5.2% of the replications, where the highest percentages correspond to 10 and 20 items and large sample sizes.

Table 3.2 shows the percentages of times AIC, BIC and HQ select SNP_1 instead of SNP_0 for scenario **A1**.

TABLE 3.2: Percentages of times AIC, BIC and HQ select SNP_1 instead of SNP_0 for scenario A1, $p = 4, 10, 20$, $n = 200, 500, 1000$

p	n	AIC	BIC	HQ
4	200	25.2%	1.8%	8%
	500	27.4%	1.8%	10.2%
	1000	28.6%	1%	10%
10	200	31.8%	8.6%	18.2%
	500	17%	2.4%	7.6%
	1000	15.4%	0.4%	2%
20	200	15%	1.2%	6.4 %
	500	6.8%	0%	1.6%
	1000	4%	0%	0.2%

Among the three information criteria considered, the AIC has the worst performance. This is evident especially with 4 and 10 items and all sample sizes and with 20 items and small sample size. In these cases, the AIC selects the SNP_1 model from 15% to 30% of times. The HQ has good performance for 10 and 20 items and large sample sizes, while with 4 items it selects the SNP_1 model around 10% of times for all sample sizes. The BIC has the best performance and only with 10 items and small sample size it selects the SNP_1 model around 9% of times.

Table 3.3 presents the empirical power of the GH_T, M_2 and LR tests for scenarios **B1**, **C1**, **D1** and **E1**.

TABLE 3.3: Empirical power of the of the GH_T, M_2 and LR tests for scenarios B1, C1, D1 and E1, $p = 4, 10, 20$, $n = 200, 500, 1000$

			$\alpha = 0.05$			$\alpha = 0.01$		
SC	p	n	GH_T	M_2	LR	GH_T	M_2	LR
B1	4	200	0.079	0.032	0.4	0.060	0	0.206
		500	0.377	0.028	0.662	0.136	0.004	0.572
		1000	0.698	0.036	0.814	0.612	0.008	0.784
	10	200	0.321	0.006	0.736	0.142	0.002	0.673
		500	0.708	0.01	0.95	0.636	0.002	0.808
		1000	0.77	0.006	1	0.764	0	0.992
	20	200	0.272	0.002	0.878	0.132	0.002	0.808
		500	0.69	0	0.992	0.55	0	0.922
		1000	0.794	0	1	0.776	0	1
C1	4	200	0.044	0.028	0.438	0.02	0.006	0.199
		500	0.644	0.044	0.772	0.352	0.01	0.566
		1000	0.894	0.044	0.926	0.826	0.006	0.868
	10	200	0.448	0.012	0.94	0.124	0.002	0.928
		500	0.968	0.016	0.964	0.96	0.002	0.964
		1000	0.966	0.01	0.984	0.966	0.002	0.984
	20	200	0.343	0.002	0.976	0.134	0	0.968
		500	0.836	0	0.976	0.612	0	0.97
		1000	0.976	0	0.978	0.956	0	0.978
D1	4	200	0.018	0.062	0.160	0.006	0.016	0.03
		500	0.18	0.08	0.288	0.06	0.008	0.115
		1000	0.336	0.07	0.418	0.258	0.014	0.231
	10	200	0.448	0.086	0.546	0.338	0.026	0.372
		500	0.696	0.108	0.805	0.648	0.026	0.653
		1000	0.788	0.092	0.945	0.778	0.024	0.851
	20	200	0.382	0.208	0.734	0.21	0.076	0.652
		500	0.748	0.172	0.9	0.664	0.034	0.848
		1000	0.834	0.168	0.926	0.806	0.052	0.904
E1	4	200	0.03	0.046	0.274	0.01	0.01	0.097
		500	0.154	0.06	0.5	0.044	0.016	0.257
		1000	0.286	0.046	0.595	0.18	0.012	0.458
	10	200	0.45	0.064	0.788	0.256	0.012	0.676
		500	0.798	0.076	0.962	0.702	0.018	0.932
		1000	0.926	0.064	1	0.888	0.014	0.994
	20	200	0.328	0.214	0.894	0.164	0.076	0.804
		500	0.786	0.18	0.98	0.68	0.07	0.908
		1000	0.878	0.23	0.998	0.868	0.082	0.992

Overall, the power of the GH_T test is high when the sample size is large, with 10 and 20 items and in particular when the true latent variable distribution largely departs from the normal one, that is under scenarios **C1** and **E1**. On the other hand, the power of the GH_T test is low with small sample sizes for all levels of α considered. With 4 items, the power of the GH_T test is low for all sample sizes when the true latent variable is distributed as a skew-normal, that is under scenarios **D1** and **E1**. Under all scenarios, especially for small sample sizes, the power of the GH_T test is higher considering a level of $\alpha = 0.05$ and is lower for $\alpha = 0.01$. For what concerns the M_2 test, it has very low or no power to detect the non-normality of the latent variable distribution under all scenarios, number of items, sample size and values of α considered. Among the three tests considered, the LR test has the highest power under all scenarios, especially with small sample sizes. However, as showed in Table 3.1, it has the worst performance in terms of Type I error rates. As observed for the GH_T test, with 4 items and all sample sizes, the power of the LR test is higher when the true latent variable is distributed as a mixture of normals than as a skew-normal. Under scenarios **C1** and **E1**, with many items and as the sample size increases, GH_T and LR have very similar power.

Table 3.4 shows the percentages of times AIC, BIC and HQ select SNP_1 instead of SNP_0 for scenarios **B1**, **C1**, **D1** and **E1**.

TABLE 3.4: Percentages of times AIC, BIC and HQ select SNP_1 instead of SNP_0 for scenarios B1, C1, D1 and E1, $p = 4, 10, 20$, $n = 200, 500, 1000$

SC	p	n	AIC	BIC	HQ
B1	4	200	62.5%	32.5%	48.7%
		500	74.9%	58.4%	67.2%
		1000	92%	77.8%	81.4%
	10	200	86.3%	69.9%	77.2%
		500	100 %	82%	96.2%
		1000	100%	99%	100%
	20	200	96.7%	82.5%	91%
		500	100%	94%	99.6%
		1000	100%	100%	100%
C1	4	200	69.3%	31.4%	51.4%
		500	87.2%	59%	78%
		1000	95.6%	85.4%	92.6%
	10	200	95%	93.8%	94.2%
		500	96.6%	96.4%	96.4%
		1000	98.4%	98.4%	98.4%
	20	200	98.6%	97%	98%
		500	98%	97%	97.6%
		1000	98%	97.8%	97.8%
D1	4	200	40.8 %	8.2%	22.2 %
		500	50%	12.4 %	30.6 %
		1000	57.6%	21.6 %	41.4%
	10	200	69%	45.2 %	57.8%
		500	91.8%	67.2%	81%
		1000	96.2%	84%	94.4%
	20	200	80.2%	69.6%	75.4%
		500	95%	85.4%	91%
		1000	95.8 %	89.8%	92.6%
E1	4	200	51 %	17.8%	33.2%
		500	61.4 %	30.2%	50.8%
		1000	67%	44.4%	59.2%
	10	200	88.2%	73 %	80.4 %
		500	99.4%	93.6 %	96.4%
		1000	100%	99.4%	100%
	20	200	97.8%	83.2 %	91.6 %
		500	99.8%	91.8 %	98.2 %
		1000	100%	99.2%	99.8%

In the majority of cases, among the three criteria, the AIC has the best performance and it has the highest percentages of selections as the SNP_1 model. This is evident especially for small sample sizes. However, the AIC selects the SNP_1 model instead of the SNP_0 model in a lot of cases also under scenario **A1**, as showed in Table 3.2. Under all scenarios, the BIC has the worst performance with small sample size especially with 4 items, where it selects the SNP_1 models only between 8% to 30% of times, depending on the scenario considered. The performance of the HQ to select the SNP_1 model under all scenarios for small sample size is in between the performance of the AIC and BIC criteria. As the sample size increases and with many items, the three criteria have the same behaviour and select the more complex model in almost the totality of cases. With 4 items, all criteria select the SNP_1 model in a higher percentage of cases when the true latent variable is distributed as a mixture of normals than as a skew-normal. Overall, from the results in Table 3.2 and Table 3.4, it is not possible to identify the criteria that has the best performance under normality and non-normality of the latent variable.

Even if not reported in Table 3.3 and Table 3.4, only in a few cases, the SNP_1 model does not reach the convergence in the optimization process, ranging from 0.2 % to 4% of the replications, where the highest percentages correspond to small sample sizes, 4 items under scenario **B1**.

3.6 Real data applications

3.6.1 American students exposure to school and neighbourhood violence

In this section we evaluate the performance of the GH_T, M_2, LR tests and the AIC, BIC and HQ criteria to detect non-normality of the the latent variable distribution on a real data set that is part of the National Longitudinal Survey of Freshman (NLSF) (for details, see http://nlsf.princeton.edu). The aim of the NLSF is to collect data to explain the minority underachievement in higher education. Data have been collected from 1999 to 2003, in four waves, to capture emergent psychological processes, measuring the degree of social integration and intellectual engagement. The survey included equal-sized samples of white, black, Asian, and Latino freshmen entering selective colleges and universities. We analyze only a part of the questionnaire that refers to the year 1999. In particular we select 9 binary items that measure violence in the neighbourhood. The items are the following:

TABLE 3.5: NLSF data: item description

Item	Question
1	In your neighborhood, before you were ten do you remember seeing homeless people on the street?
2	Prostitutes on street?
3	Gang members hanging out on the street?
4	Drug paraphernalia on the street?
5	People selling illegal drugs in public?
6	People using illegal drugs in public?
7	People drinking or drunk in public?
8	Physical violence in public?
9	Hearing the sound of gunshots?

The original sample of observations is composed by 3924 observations . Possible responses are “no”, “yes”, “don’t know” and “refused”. Individuals that have responded “don’t know” or “refused” are excluded from the analysis, while responses “no” are coded as 0 and “yes” as 1. The data set is finally composed by 3891 observations. A sub-sample of 400 observations and a larger number of items of the questionnaire that refers to the year 1999 has been analyzed also by Cagnone and Viroli (2012), that fitted a latent trait models with two factors distributed as a mixture of normals. They found that the item reported in Table 3.5 are highly loaded only on one factor, distributed as a mixture of normals. Some preliminary descriptive analysis reported in this section and goodness-of-fit measures for the SNP_0 model are performed with the “ltm” R package.

Table 3.6 reports the proportion of “0” (no) and “1” (yes) responses in each item.

TABLE 3.6: NLSF data: proportion of “0” (no) and “1” (yes) responses in each item

Item	0	1
1	0.76	0.23
2	0.96	0.04
3	0.85	0.15
4	0.89	0.11
5	0.92	0.08
6	0.93	0.06
7	0.69	0.31
8	0.75	0.25
9	0.85	0.15

In all the items the proportion of “0” responses is higher than “1” and, even if not reported in the table results, all the items are significantly associated among each other. The first step of the analysis is to fit the SNP_0 model to the data.

Table 3.7 reports the value of the Pearson’s chi-square X^2 test and the associated p-value for the SNP_0 model.

TABLE 3.7: NLSF data: Pearson’s chi-square test and associated p-value for the SNP_0 model

X^2 statistic	Degrees of freedom	p-value
682.30	212	0

According to the Pearson’s chi-square test, the SNP_0 model does not have a good fit to the data. Since the data are sparse and 174 observed response patterns out of the total 231 observed response patterns have expected frequencies less than 5, we should inspect the residuals calculated from marginal frequencies (Bartholomew et al., 2011).

Table 3.8 reports the chi-squared residuals calculated from the two-way margins for the SNP_0 model.

TABLE 3.8: NLSF data: two-way margins residuals for the SNP_0 model

Response	Item i	Item j	Obs	Exp	$(O-E)^2/E$
(0,0)	7	8	2405	2370.88	0.49
	1	2	2949	2921.59	0.26
	1	8	2483	2506.78	0.23
(1,0)	4	7	74	58.53	4.09 ***
	5	7	36	26.52	3.39
	6	7	17	26.36	3.32
(0,1)	1	2	16	43.19	17.12 ***
	1	5	91	72.22	4.88 ***
	1	6	79	62.19	4.54 ***
(1,1)	1	2	144	117.84	5.81 ***
	2	6	66	78.56	2.01
	4	5	239	218.46	1.93

Note 1:'***' denotes a chi-squared residual greater than 4

We consider the rule of thumb that residuals greater than 4 are indicators of bad fit of the correspondent pair of items. On these data, the SNP_0 model does not have a good fit for some pairs of items.

Table 3.9 reports the chi-squared residuals calculated from the three-way margins for the SNP_0 model.

TABLE 3.9: NLSF data: three-way margins residuals for the SNP_0 model

Response	Item i	Item j	Item k	Obs	Exp	$(O-E)^2/E$
(0,0,0)	6	7	8	2396	2358.91	0.58
	1	2	3	2757	2718.45	0.55
	1	8	9	2332	2366.54	0.50
(1,0,0)	4	7	8	47	27.51	13.81 ***
	5	7	8	20	9.59	11.31 ***
	4	6	7	68	50.59	5.99 ***
(0,1,0)	1	2	4	4	27.34	19.92 ***
	1	2	5	9	31.04	15.65 ***
	1	2	6	11	33.63	15.23 ***
(1,1,0)	1	2	6	83	48.85	23.88 ***
	1	2	4	54	29.30	20.83 ***
	1	2	5	64	36.54	20.64 ***
(0,0,1)	1	3	6	47	29.50	10.38 ***
	1	2	5	84	60.07	9.53 ***
	1	2	6	74	52.63	8.68 ***
(1,0,1)	4	5	8	116	143.56	5.29 ***
	1	4	5	44	61.96	5.21 ***
	4	5	7	144	173.83	5.12 ***
(0,1,1)	3	6	7	84	56.48	13.41 ***
	1	4	5	58	37.34	11.42 ***
	1	2	8	9	25.07	10.30 ***
(1,1,1)	1	2	7	128	108.87	3.36
	2	6	9	48	61.38	2.91
	2	3	6	55	69.18	2.91

Note 1:'***' denotes a chi-squared residual greater than 4

According to the same rule on the residuals, the SNP_0 model does not have a good fit also for some triplets of items.

It is worth computing the M_2 test statistic of Maydeu-Olivares and Joe (2005) that, as described in section 3.3, is not affect by the problem of sparse data.

Table 3.10 reports the value of the M_2 test statistic and the associated p-value for the SNP_0 model.

TABLE 3.10: NLSF data: M_2 test and associated p-value for the SNP_0 model

M_2 statistic	Degrees of freedom	p-value
182.33	27	0

Also according to the M_2 test, the SNP_0 model does not have a good fit to the data for $\alpha = 0.05$. However, the M_2 test does not reveal the source of misfit. It could be the non-normality of the latent variable even if, as showed in the simulation study, the M_2 test has a very low power to detect this type of model misspecification, or other types of model misspecification.

In order to evaluate if the misfit of the SNP_0 model is due to the non-normality of the distribution of the latent variable, we estimate the SNP_1 model.

Table 3.11 reports parameter estimates and related standard errors based on the SNP_0 and SNP_1 models.

TABLE 3.11: NLSF data: SNP_0 and SNP_1 parameter estimates and related standard errors

Parameter (SD)	SNP0	SNP1
α_{01}	-1.85 (0.07)	-2.61 (0.18)
α_{02}	-5.21 (0.25)	-6.56 (0.38)
α_{03}	-3.43 (0.14)	-4.88 (0.26)
α_{04}	-5.05 (0.26)	-7.49 (0.46)
α_{05}	-6.85 (0.47)	-9.43 (0.60)
α_{06}	-5.75 (0.31)	-7.76 (0.45)
α_{07}	-1.61 (0.09)	-2.78 (0.24)
α_{08}	-1.99 (0.08)	-2.99 (0.21)
α_{09}	-2.76 (0.10)	-3.68 (0.19)
α_{11}	1.94 (0.09)	2.93 (0.20)
α_{12}	2.39 (0.17)	4.09 (0.32)
α_{13}	2.84 (0.15)	4.64 (0.27)
α_{14}	3.77 (0.23)	6.71 (0.46)
α_{15}	4.56 (0.36)	7.75 (0.55)
α_{16}	3.38 (0.23)	5.83 (0.39)
α_{17}	2.77 (0.14)	4.16 (0.30)
α_{18}	2.32 (0.11)	3.58 (0.24)
α_{19}	2.00 (0.10)	3.19 (0.20)
φ_1	-	0.23(0.04)

Parameter estimates are quite different among the two methods. Since the SNP_0

and the SNP_1 models differ only in the φ_1 parameter, we compute the LR test for nested models (Table 3.12).

TABLE 3.12: NLSF data: the LR test and the associated p-value

LR statistic	Degrees of freedom	p-value
7.65	1	0.006

According to the LR test, the SNP_1 model has a better fit to the data. However, as showed in the simulation study, with many items the performance of the LR test is not very good in terms of Type I error rates, also for large sample sizes. For this reason, it is also worth computing the information criteria and the GH_T test.

Table 3.13 reports the values of the AIC, BIC and HQ criteria for the SNP_0 and SNP_1 models.

TABLE 3.13: NLSF data: Information criteria for the SNP_0 and SNP_1 models

	AIC	BIC	HQ
SNP_0	21309.32	21422.12	21349.36
SNP_1	21303.67	21422.73	21345.93

The AIC and HQ criteria select the SNP_1 model, while the BIC criterion the SNP_0 model. These results do not give a clear indication on the normality or non-normality of the latent variable.

Table 3.14 reports the value of the GH_T test statistic and the associated p-value.

TABLE 3.14: NLSF data: the GH_T test and the associated p-value

GH_T statistic	Degrees of freedom	p-value
245.68	3.45	0

Since the p-value associated to the GH_T test is less than the nominal level $\alpha = 0.05$, we reject the null hypothesis that the latent variable is normally distributed. This result is coherent with the simulation results, where the GH_T test shows good performance in terms of power to detect non-normality of the latent variable distribution with many items and large sample sizes. Moreover, since in the simulation study the test GH_T has good performance in terms of Type I error rates with many items, we could be confident in rejecting the null hypothesis of normality of the latent variable. In this analysis, the GH_T turns out to be very useful to detect non-normality of the latent variable distribution because the information criteria show contradictory results.

3.6.2 Data set on sex role expectations

In this section we evaluate the performance of the GH_T, M_2, LR tests and the AIC, BIC and HQ criteria to detect non-normality of the the latent variable distribution on the real data set on sex role expectations analyzed in section 2.7. The data set is composed by 257 observations and 4 items and the detailed data description can be found in section 2.7. Also in this example we report some preliminary descriptive analysis.

Table 3.15 reports the proportion of "0" (no) and "1" (yes) responses in each item.

TABLE 3.15: Data on sex role expectations: proportion of "0" (no) and "1" (yes) responses in each item

Item	0	1
1	0.66	0.34
2	0.70	0.30
3	0.67	0.33
4	0.55	0.45

We can notice that the proportion of "0" is higher than "1" in all the items. Even if not reported in the table results, there is a significant association among all items. We first fit the SNP_0 model to the data.

Table 3.16 reports the value of the Pearson's chi-square X^2 test and the associated p-value for the SNP_0 model.

TABLE 3.16: Data on sex role expectations: Pearson's chi-square test and associated p-value for the SNP_0 model

X^2 statistic	Degrees of freedom	p-value
23.29	7	0

Given that the p-value of the X^2 test is less than 0.05, we reject the null hypothesis that the SNP_0 model holds. However, also in this case the data are sparse and some of the expected cells frequencies are small. Thus, as in the previous example, we consider the chi-squared marginal residuals.

Table 3.17 reports the chi-squared residuals calculated from the two-way margins for the SNP_0 model.

TABLE 3.17: Data on sex role expectations: two-way margins residuals for the SNP_0 model

Response	Item i	Item j	Obs	Exp	$(O-E)^2/E$
(0,0)	1	2	138	130.54	0.43
	3	4	126	123.54	0.05
	2	4	115	117.24	0.04
(1,0)	1	2	41	47.93	1.00
	2	4	26	22.76	0.46
	1	3	41	38.29	0.19
(0,1)	1	2	32	39.13	1.30
	2	4	64	61.24	0.12
	1	3	40	38.87	0.03
(1,1)	1	2	46	39.40	1.11
	2	4	52	55.76	0.25
	1	3	46	49.04	0.19

All the residuals are small, meaning that the SNP_0 model has a good fit for all pair of items. However, it is advisable to inspect also higher order residuals (Table 3.18).

TABLE 3.18: Data on sex role expectations: three-way margins residuals for the SNP_0 model

Response	Item i	Item j	Item k	Obs	Exp	$(O-E)^2/E$
(0,0,0)	1	3	4	98	100.12	0.04
	2	3	4	106	107.61	0.02
	1	2	3	110	111.37	0.02
(1,0,0)	2	3	4	20	15.94	1.04
	1	3	4	28	23.42	0.89
	1	2	3	32	29.28	0.25
(0,1,0)	1	3	4	11	9.25	0.33
	2	3	4	9	9.64	0.04
	1	2	4	16	15.22	0.04
(1,1,0)	1	3	4	4	7.21	1.43
	1	2	4	10	7.54	0.80
	2	3	4	6	6.83	0.10
(0,0,1)	1	2	3	28	19.17	4.07 ***
	1	2	4	45	36.40	2.03
	2	3	4	36	33.05	0.26
(1,0,1)	1	2	3	9	18.65	4.99 ***
	1	2	4	19	24.84	1.37
	2	3	4	9	12.50	0.98
(0,1,1)	1	2	3	12	19.70	3.01
	1	2	4	16	23.90	2.61
	1	3	4	29	29.61	0.01
(1,1,1)	1	2	3	37	30.39	1.44
	1	2	4	36	31.86	0.54
	2	3	4	43	43.26	0.00

Note 1:'***' denotes a chi-squared residual greater than 4

According to the three-way margins residuals, the SNP_0 model does not have a good fit for the triplet of items 1, 2 and 3.

Table 3.19 reports the M_2 test statistic and the associated p-value for the SNP_0 model.

TABLE 3.19: Data on sex role expectations: M_2 test statistic and associated p-value for the SNP_0 model

M_2 statistic	Degrees of freedom	p-value
8.59	2	0.01

The M_2 test rejects the null hypothesis that the SNP_0 model holds for $\alpha = 0.05$ but it does not give any information on the type of model misspecification. As in the previous example, it could be the non-normality of the latent variable or others sources of misfit as, for example, the presence of some aberrant responses for the SNP_0 model (found on these data by Mavridis and Moustaki, 2009).

To assess the non-normality of the distribution of the latent variable, we consider the SNP_1 model.

Table 3.20 reports parameter estimates and related standard errors based on the SNP_0 and SNP_1 models.

TABLE 3.20: Data on sex role expectations: SNP_0 and SNP_1 parameter estimates and related standard errors

Parameter	SNP0	SNP1
α_{01}	-0.82 (0.17)	-0.75 (0.16)
α_{02}	-1.15 (0.21)	-0.99 (0.19)
α_{03}	-1.93 (0.75)	-0.98 (0.64)
α_{04}	-0.32 (0.22)	0.11 (0.25)
α_{11}	1.09 (0.23)	1.21 (0.28)
α_{12}	1.51 (0.30)	1.62 (0.39)
α_{13}	4.36 (1.71)	6.46 (5.23)
α_{14}	2.27 (0.50)	3.00 (0.91)
φ_1	-	-0.54 (0.07)

Note 1: Standard errors in round brackets

Differences among parameter estimates computed with the two models are evident especially for the intercepts and slopes of the items 3 and 4. The slope of the item 3 has a very large standard error, especially under the SNP_1 model.

As in the previous data example, we inspect the value of the LR test statistic and the associated p-value, reported in Table 3.21.

TABLE 3.21: Data on sex role expectations: the LR test and the associated p-value

LR statistic	Degrees of freedom	p-value
10.55	1	0.001

According to the LR test, we reject the null hypothesis of normality of the latent variable distribution. In the simulation study, the LR test has the highest power to detect non-normality of the latent variable distribution with few items and small sample sizes. However, with 4 items, the LR test has also seriously inflated false positive rates under some sample sizes. For this reason, we cannot rely on the result of this test.

Table 3.22 reports the values of the AIC, BIC and HQ criteria for the SNP_0 and SNP_1 models.

TABLE 3.22: Data on sex role expectations: Information criteria for the SNP_0 and SNP_1 models

	AIC	BIC	HQ
SNP_0	1188.14	1216.54	1199.55
SNP_1	1179.59	1211.53	1192.43

According to all criteria, the SNP_1 model has a better fit to the data than the SNP_0 model. However, in the simulation results, the information criteria do not show good performance with small sample size and a few items. In particular the AIC and HQ select the more complex model in a lot of cases under normality of the latent variable and the BIC has low power to select the SNP_1 model under non-normality of the latent variable. For these reasons, we cannot rely completely on the results given by the information criteria.

Table 3.23 reports the values of the GH_T test and the associated p-value.

TABLE 3.23: Data on sex role expectations: the GH_T test and the associated p-value

GH_T statistic	Degrees of freedom	p-value
0.57	1.03	0.46

Since the p-value associated to the GH_T test is greater than the nominal levels $\alpha = 0.05$, we do not reject the null hypothesis that the latent variable is normally distributed. In the simulation, as observed for the information criteria, the GH_T test does not have good performance for small sample size and a few items, especially in terms of power. Moreover, the large variance of the slope of the item 3 negatively affects the performance of the GH_T test. We cannot neither rely on the result of the GH_T test.

For this small data set, it can be useful to inspect the chi-squared residuals for the SNP_1 model.

Table 3.24 reports the chi-squared residuals calculated from the two-way margins for the SNP_1 model.

TABLE 3.24: Data on sex role expectations: two-way margins residuals for the SNP_1 model

Response	Item i	Item j	Obs	Exp	$(O-E)^2/E$
(0,0)	1	2	138	133.72	0.14
	2	4	115	118.55	0.11
	1	3	130	133.26	0.08
(1,0)	2	4	26	22.29	0.62
	1	3	41	37.18	0.39
	1	2	41	45.13	0.38
(0,1)	1	2	32	36.21	0.49
	1	3	40	36.66	0.30
	2	4	64	60.30	0.23
(1,1)	1	2	46	41.94	0.39
	1	3	46	49.88	0.30
	2	4	52	55.87	0.27

As observed for the SNP_0 model, the SNP_1 model has a good fit for all pairs of items. As before, we inspect also higher order residuals.

Table 3.25 reports the chi-squared residuals calculated from the three-way margins for the SNP_1 model.

TABLE 3.25: Data on sex role expectations: three-way margins residuals for the SNP_1 model

Response	Item i	Item j	Item k	Obs	Exp	$(O-E)^2/E$
(0,0,0)	1	3	4	98	100.86	0.08
	1	2	3	110	112.94	0.08
	1	2	4	93	94.783	0.03
(1,0,0)	1	3	4	28	24.02	0.66
	3	3	4	20	17.02	0.52
	1	2	3	32	29.66	0.18
(0,1,0)	2	3	4	9	10.69	0.27
	1	3	4	11	10.10	0.08
	1	2	3	20	20.32	0.00
(1,1,0)	1	2	4	10	6.11	2.47
	1	3	4	4	5.85	0.58
	1	2	3	9	7.52	0.29
(0,0,1)	1	2	3	28	20.78	2.51
	1	2	4	45	38.94	0.94
	2	3	4	36	34.74	0.04
(1,0,1)	1	2	3	9	15.47	2.70
	2	3	4	9	10.82	0.31
	1	2	4	19	21.36	0.26
(0,1,1)	1	2	3	12	15.88	0.95
	1	2	4	16	20.03	0.81
	2	3	4	28	25.56	0.24
(1,1,1)	1	2	3	37	34.42	0.19
	1	3	4	42	44.03	0.09
	2	3	4	43	45.04	0.09

All residuals are less than 4, indicating that the SNP_1 model has good fit for all triplets of items. Compared to the SNP_0 model, the SNP_1 model has a better local fit to the data. However, assessing the normality and the non-normality of the latent variable on small data sets remains an open issue.

3.7 Discussion

In this work, we extended the use of the Generalized Hausman test to detect non-normality of the latent variable distribution by considering the SNP-IRT model. We evaluated the performance of the GH_T test by means of a simulation study and in real data analysis and we compared it with the M_2 and the LR tests. Moreover, we computed the classic AIC, BIC and HQ criteria.

The simulation study highlights that, when the latent variable is normally distributed, the GH_T test has good performance in terms of Type I error rates with many items and in general for large sample sizes. For what concerns the power, the GH_T test has good performance to detect non-normality of the latent variable especially when the true latent variable largely departs from a normal distribution, with many items and large sample sizes. However, the power of the GH_T test is dramatically affected by small sample sizes and a few number of items. The M_2 test has good performance in terms of Type I error rates but it has very low or no power to detect the non-normality of the latent variable. On the contrary, the LR test has the highest power especially for small sample sizes but, among the three test considered, it has the worst performance in terms of Type I error rates. From the simulation results it is not possible to identify the best information criterion under normality and non-normality of the latent variable. Indeed, the AIC tends to select the more complex model in the highest percentages of cases under non-normality and normality of the latent variable. The BIC has the best performance under normality of the latent variable but the worst under non-normality for small sample sizes. The performance of the HQ criterion is in between the one of the AIC and BIC. In general, all criteria worsen their performance with a few items and small sample sizes, both under normality and non-normality of the latent variable. In the example on the NLSF data set, the GH_T test turns out to be very useful to detect the non-normality of the latent variable because the information criteria show contradictory results. In the application on the small data set on sex role expectations, none of the methods considered are useful to assess the normality and non-normality of the latent variable. In this case, chi-squared marginal residuals may help to assess the local fit of the models considered. However, detecting non-normality of the latent variable on small data sets remains an open issue and, for further research, different methods to overcome the limits of the GH_T test and the information criteria could be explored.

Moreover, for further studies on the performance of the Generalized Hausman test in the IRT context, other types of model violations could be considered, as local dependence or violation of the item characteristic function, as it has been done by Ranger and Much (2020) for the classic Hausman test. In these cases, other types of estimation methods consistent under model misspecification should be considered in order to apply the test.

Another possible extension for further research could be to explore the performance of the GH and GH_T tests to select the number of classes in latent class models for binary data. Something similar has been done by Bartolucci et al. (2017), who derived a robust version of the Hausman test for GLMM, where a discrete distribution for the random effects is assumed. A latent class model can be seen as a GLMM based on finite-mixture formulation, that is a discrete distribution for the random effects/ latent variables, without the inclusion of covariates. Analogies and differences among the two formulations can be found in Bartolucci et al. (2017). To implement the GH and GH_T tests, the Conditional Maximum Likelihood (CML) estimation method (Tchetgen and Coull, 2006) could be considered because it gives consistent estimates of the fixed covariate effects in the GLMM formulation and of the difficulty parameters in IRT models even under incorrect assumptions of the distribution of the random effects/latent variables (Bartolucci et al., 2017). The GH and GH_T tests should be based on the difference between the difficulty parameters obtained with the CML method and those obtained with the classic Marginal Maximum Likelihood estimation method, where each time a different number of latent classes is assumed in the model. As suggested by Bartolucci et al. (2017), a sequential strategy consisting in increasing the number of latent classes could be adopted: the correct number of classes is identified as the first one for which the test does not reject the null hypothesis. The performance of the GH and GH_T tests could be compared with the classic information criteria, that are usually adopted in latent class models to select the number of classes.

Chapter 4

Conclusions

This thesis covered the problem of assessing the fit of unidimensional IRT models for binary data under different types of model misspecification. We considered some robust tests derived by White (1982), the generalized Lagrange Multiplier (LM(S)) test for the local fit and the generalized Hausman (GH) test for the global fit of the model, in the IRT context.

In particular, we assessed measurement invariance through the LM(S) test, whose expression involves the sandwich covariance matrix of the score function, under two types of IRT model misspecification, local dependence among items and non-normality of the latent variable distribution. We considered in the simulations and in the real data analysis the so-called Multiple Indicator Multiple Causes model, where a binary group covariate affected both the item(s) and the latent variable. The performance of the LM(S) test was compared with other versions of the Lagrange Multiplier test, that differ in the form of the covariance matrix of the score function. They are the classical Hessian (LM(H)) and cross-product (LM(CP)) Lagrange Multiplier tests and the Jackknife score (GS(J)) test, derived under model misspecification. The power of the LM(S), LM(H) and LM(CP) tests was computed empirically and asymptotically.

The following Table summarizes some of the most important results of Chapter 2, reported in section 2.5. This Table shows the test(s) that has (have) the best performance for each model misspecification, scenario (SC), number of items p and sample size n. The scenarios **A**,**B**,**C** were considered to study the false positive rates and **D**,**E**,**F** the power of the tests. Different levels of local dependence in the data generating models were given by the percentage of local dependent items (% of LD) and the variance of the random effects u (σ_u^2). Non-normality of the latent variable distribution was introduced in the data generating models by considering a mixture of normals and different skew-normal (SN) distributions for the latent variable ϵ.

TABLE 4.1: Summary results of Chapter 2

Misspecification	SC	p	n	Test
20,50% of LD, $\sigma_u^2 = 0.25$	**A**,**B**,**C**	10,20	200,500,1000,5000	LM(H),LM(S)
20,50% of LD, $\sigma_u^2 = 1, 2.25$	**A**,**B**,**C**	10,20	200,500,1000,5000	None
20,50% of LD, $\sigma_u^2 = 0.25, 1, 2.25$	**D**,**E**,**F**	10,20	200	LM(CP),LM(H)
	D,**E**,**F**	10,20	500,1000,5000	LM(CP),LM(H),LM(S)
$\epsilon \sim 0.3N(-1.5, 0.2) + 0.7N(1, 0.4)$	**A**,**B**,**C**	10,20	200,500,1000	LM(S),LM(H)
	D,**E**,**F**	10,20	200,500,1000	LM(CP),LM(H),LM(S)
$\epsilon \sim SN(1)$	**A**,**B**,**C**	10,20	200,500,1000	LM(S),LM(H)
	D,**E**,**F**	10,20	200,500,1000	LM(CP)
$\epsilon \sim SN(3)$	**A**,**B**,**C**	10,20	200,500,1000	None
	D,**E**,**F**	10,20	200	LM(CP)
	D,**E**,**F**	10,20	500,1000	LM(H),LM(CP)

From Table 4.1 it is evident that none of the tests has overall superior performance compared with others. In general, when the misspecification is low, all tests have good performance in terms of false positive rates and power for large sample sizes. When the misspecification is high, the test performance deteriorates. In particular, the LM(CP) test has the lowest performance in terms of false positive rates under all kinds of misspecification. The power of the LM(H) and LM(S) tests is affected by the non-normality of the latent variable distribution, especially for small sample sizes. This is evident also in the real data example on sex role expectations.

As regards the asymptotic power, between the two methods used to compute it, described in sections 2.3.1 and 2.3.2, we recommend the first one for the simpler calculation of the non-centrality parameter. In general the asymptotic power of the tests computed with the two methods turns out to be very close to the empirical one. Finally, the GS(J) test has never a better performance than the LM(S) test and it is computationally intense.

We then focused on the non-normality of the latent variable distribution and we extended the GH test to detect this type of model misspecification. The GH test was obtained as the difference of the estimators of the classic IRT model for binary data and the seminonparametric (SNP)-IRT model, that allows for a more flexible shape of the latent variable distribution. To avoid the inversion of the covariance matrix of the difference of the parameter estimates, we considered an alternative form of this test, that we called GH_T test, both in the simulations and in the real data analysis. The performance of the GH_T test was compared with two overall goodness-of fit tests, the M_2 test, a limited information statistic not affected by the problem of sparse data, and the classic Likelihood-Ratio (LR) test for nested models. Information criteria (IC) were also computed. The next Table summarizes the main results obtained in Chapter 3. This Table reports the test(s) that has (have) the best performance in terms of Type I error rates and power for each scenario (SC), number of items p and sample size n. The scenario **A1** was considered to study the Type I error rates of the tests. As for the power analysis, we considered different degrees of mixture of two normals (scenarios **B1**,**C1**) and skew-normals (scenarios **D1** and **E1**) as the true distribution of the latent variable. Table 4.2 shows also the information criterion (criteria) that has (have) the best performance to select the correct model, in percentages of times, under the different simulation conditions.

TABLE 4.2: Summary results of Chapter 3

SC	p	n	Test/IC
A1	4	200,500	M_2/BIC
	4	1000	M_2,GH_T/BIC
	10,20	200,500,1000	M_2,GH_T/BIC
B1,**C1**,**D1**,**E1**	4,10,20	200	LR/AIC
B1,**D1**	4,10,20	500,1000	LR/AIC,HQ
C1	4,10,20	500,1000	LR,GH_T/AIC,HQ,BIC
E1	4	500,1000	LR/AIC,HQ
	10,20	500,1000	LR,GH_T/AIC,HQ,BIC

Considering the Type I error rates and power simulation results and the analysis on the NLSF dataset, the GH_T test has the best performance for many items and in general for large sample sizes. The M_2 test has the worst performance in terms of power and the LR test in terms of Type I error rates. It is not possible to identify the IC that has the best performance both under normality and non-normality of the latent variable distribution.

A line of future research can be to assess the fit of IRT models for polytomous data under model misspecification. Polytomous items include more than two categories that can be ordered or unordered. The principal IRT models used to analyze these data are the Nominal Response model (Bock, 1972), the Graded Response model (Samejima, 1969) and the Partial Credit model (Masters, 1982). Polytomous items provide more information than binary items concerning the level of the latent trait (Samejima, 1969, Donoghue, 1994). This can result in more accurate parameter estimates and related standard errors. Moreover, in general test statistics, as the limited information ones, are more powerful for polytomous than binary items (Maydeu-Olivares, 2013). The performance of the LM(S) test could be studied for this types of items, under the different types of model violations considered in Chapter 2. Moreover, the GH and GH_T tests presented in Chapter 3 could be applied to IRT models for polytomous data, assuming the SNP representation of the latent variable distribution, to detect non-normality of the latent variable distribution. Since these models involve a higher number of parameters, the additional issue, compared to binary data, could be the computational cost of the estimation process (Bartholomew et al., 2011). With polytomous items we expect an improvement in the performance of the LM(S), GH and GH_T tests, especially in terms of power.

Appendix A

Appendix

A.1 Derivation of the Generalized Lagrange Multiplier Test

To derive the LM(S) test, we use the same procedure of Boos (1992) and Bera et al. (2020). We consider the set of hypotheses (2.10) and the same notation of section 2.3.2.

Consider the Taylor expansion of $S_1(\tilde{\boldsymbol{\theta}}_n)$

$$0 = \sqrt{n}S_1(\tilde{\boldsymbol{\theta}}_n) = \sqrt{n}S_1(\boldsymbol{\theta}_*) + \sqrt{n}\frac{\partial S_1(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_{*1}}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) + \sqrt{n}\frac{\partial S_1(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_{*2}}(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{*2}), \quad \text{(A.1)}$$

and $S_2(\tilde{\boldsymbol{\theta}}_n)$

$$\sqrt{n}S_2(\tilde{\boldsymbol{\theta}}_n) = \sqrt{n}S_2(\boldsymbol{\theta}_*) + \sqrt{n}\frac{\partial S_2(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_{*1}}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) + \sqrt{n}\frac{\partial S_2(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_{*2}}(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{*2}), \quad \text{(A.2)}$$

where $S_1(\tilde{\boldsymbol{\theta}}_n)$ is the subset of $S(\tilde{\boldsymbol{\theta}}_n)$ that corresponds to the parameters $\boldsymbol{\theta}_{*1}$ evaluated at $\tilde{\boldsymbol{\theta}}_n$ and $S_2(\tilde{\boldsymbol{\theta}}_n)$ is defined as in section 2.3.2. $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_2$ denote the subsets of $\tilde{\boldsymbol{\theta}}_n$ corresponding to the parameter $\boldsymbol{\theta}_{*1}$ and $\boldsymbol{\theta}_{*2}$, respectively.

If H_0 holds, $\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{*2}$ is $\mathbf{0}$ because $\boldsymbol{\theta}_{*2} = \mathbf{c}$ and $\tilde{\boldsymbol{\theta}}_2 = \mathbf{c}$. Now $\frac{\partial S \boldsymbol{\theta}_*}{\partial \boldsymbol{\theta}_*} = -\hat{A}(\boldsymbol{\theta}_*)$, and replacing $\hat{A}(\boldsymbol{\theta}_*)$ with its asymptotically expected version $A(\boldsymbol{\theta}_*)$ we obtain from equation (A.1)

$$0 = \sqrt{n}S_1(\boldsymbol{\theta}_*) - \sqrt{n}A_{11}(\boldsymbol{\theta}_*)(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) \quad \text{(A.3)}$$

that becomes

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) = \sqrt{n}A_{11}(\boldsymbol{\theta}_*)^{-1}S_1(\boldsymbol{\theta}_*) \quad \text{(A.4)}$$

and from equation (A.2)

$$\sqrt{n}S_2(\tilde{\boldsymbol{\theta}}_n) = \sqrt{n}S_2(\boldsymbol{\theta}_*) - \sqrt{n}A_{21}(\boldsymbol{\theta}_*)(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}). \tag{A.5}$$

Substituting (A.4) in (A.5) we obtain

$$\begin{aligned}\sqrt{n}S_2(\tilde{\boldsymbol{\theta}}_n) &= \sqrt{n}S_2(\boldsymbol{\theta}_*) - \sqrt{n}A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}S_1(\boldsymbol{\theta}_*) \\ &= [I_2, -A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}]\begin{bmatrix}\sqrt{n}S_2(\boldsymbol{\theta}_*) \\ \sqrt{n}S_1(\boldsymbol{\theta}_*)\end{bmatrix},\end{aligned} \tag{A.6}$$

where I_2 is the identity matrix of the same dimension of S_2.

Under H_0 the expected value of $S_2(\tilde{\boldsymbol{\theta}}_n)$ is 0 and the $V(S_2(\tilde{\boldsymbol{\theta}}_n)) = E[S_2(\tilde{\boldsymbol{\theta}}_n)S_2^T(\tilde{\boldsymbol{\theta}}_n)]$. Therefore, the asymptotic covariance matrix of $S_2(\tilde{\boldsymbol{\theta}}_n)$ is (Boos, 1992)

$$\begin{aligned}V(S_2(\boldsymbol{\theta}_*)) &= [I_2, -A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}]B(\boldsymbol{\theta}_*)[I_2, -A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}]' \\ &= B_{22}(\boldsymbol{\theta}_*) - A_{21}(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)B_{12}(\boldsymbol{\theta}_*) - B_{21}(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)A_{12}(\boldsymbol{\theta}_*) + \\ &+ A_{21}(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)B_{11}(\boldsymbol{\theta}_*)(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)A_{12}(\boldsymbol{\theta}_*)\end{aligned} \tag{A.7}$$

Through matrix manipulation it is possible to show that equation (A.7) is equivalent to (Boos and Stefanski, 2013)

$$V(S_2(\boldsymbol{\theta}_*)) = A^{22}(\boldsymbol{\theta}_*)C_{22}(\boldsymbol{\theta}_*)A^{22}(\boldsymbol{\theta}_*), \tag{A.8}$$

where the matrix A^{22} is computed according to formula (2.6) and evaluated at $\boldsymbol{\theta}_*$. The asymptotic distribution of $S_2(\tilde{\boldsymbol{\theta}})$ under H_0 is

$$S_2(\tilde{\boldsymbol{\theta}}_n) \sim N(0, V(S_2(\boldsymbol{\theta}_*))) \tag{A.9}$$

An estimator of (A.9) is its observed version of evaluated at $\tilde{\boldsymbol{\theta}}_n$

$$V(S_2(\tilde{\boldsymbol{\theta}}_n)) = \hat{A}^{22}(\tilde{\boldsymbol{\theta}}_n)\hat{C}_{22}(\tilde{\boldsymbol{\theta}}_n)\hat{A}^{22}(\tilde{\boldsymbol{\theta}}_n). \tag{A.10}$$

The LM(S) test statistic, constructed as $S'V^{-1}S$ (Boos, 1992), is

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

Under H_0 the statistics LM(S) is asymptotically distributed as a χ^2 with degrees of freedom equal to the number of constrained parameters under H_0.

A.2 Derivation of the non-centrality parameter of the Generalized Lagrange Multiplier Test

To obtain the non-centrality parameter of the LM(S) test, let's consider the distribution of the LM(S) test under a sequence of local alternative around H_0. The hypotheses H_0 and H_1 can be formalized as follows:

$$H_0 : \boldsymbol{\theta}'_{*2} = \mathbf{c} \quad vs \quad H_1 : \boldsymbol{\theta}'_{*2} = \mathbf{c} + \frac{\boldsymbol{\xi}}{\sqrt{n}}, \tag{A.12}$$

where $\mathbf{c}$ is a vector of constants and $\boldsymbol{\xi}$ is an arbitrary vector of dimension $\boldsymbol{\theta}_{*2}$.

If H_1 holds and we replace the matrix $\hat{A}$ with its asymptotic expected version A, equation (A.1) becomes

$$\begin{aligned} 0 &= \sqrt{n}S_1(\boldsymbol{\theta}_*) - \sqrt{n}A_{11}(\boldsymbol{\theta}_*)(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) + \sqrt{n}\frac{\partial S_1(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_{*2}}(\mathbf{c} - \mathbf{c} - \frac{\boldsymbol{\xi}}{\sqrt{n}}) \\ &= \sqrt{n}S_1(\boldsymbol{\theta}_*) - \sqrt{n}A_{11}(\boldsymbol{\theta}_*)(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) + \sqrt{n}A_{12}(\boldsymbol{\theta}_*)(\frac{\boldsymbol{\xi}}{\sqrt{n}}) \end{aligned} \tag{A.13}$$

and we get

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) = \sqrt{n}A_{11}(\boldsymbol{\theta}_*)^{-1}S_1(\boldsymbol{\theta}_*) - A_{11}(\boldsymbol{\theta}_*)^{-1}A_{12}(\boldsymbol{\theta}_*)\boldsymbol{\xi} \tag{A.14}$$

and equation (A.2) becomes

$$\begin{aligned} \sqrt{n}S_2(\tilde{\boldsymbol{\theta}}_n) =& \sqrt{n}S_2(\boldsymbol{\theta}_*) + \sqrt{n}\frac{\partial S_2(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_{*1}}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) + \sqrt{n}\frac{\partial S_2(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_{*2}}(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{*2}) \\ =& \sqrt{n}S_2(\boldsymbol{\theta}_*) - \sqrt{n}A_{21}(\boldsymbol{\theta}_*)(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{*1}) + A_{22}(\boldsymbol{\theta}_*)\boldsymbol{\xi}. \end{aligned} \tag{A.15}$$

.

Substituting (A.14) in (A.15) we obtain

$$\begin{aligned}\sqrt{n}S_2(\tilde{\boldsymbol{\theta}}_n) =& \sqrt{n}S_2(\boldsymbol{\theta}_*) - \sqrt{n}A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}S_1(\boldsymbol{\theta}_*) - A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}A_{12}(\boldsymbol{\theta}_*)\boldsymbol{\xi} + A_{22}(\boldsymbol{\theta}_*)\boldsymbol{\xi}\\ =& \sqrt{n}S_2(\boldsymbol{\theta}_*) - \sqrt{n}A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}S_1(\boldsymbol{\theta}_*) + [A_{22}(\boldsymbol{\theta}_*) - A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}A_{12}(\boldsymbol{\theta}_*)]\boldsymbol{\xi}\\ =& [I_2, -A_{21}(\boldsymbol{\theta}_*)A_{11}(\boldsymbol{\theta}_*)^{-1}]\begin{bmatrix}\sqrt{n}S_2(\boldsymbol{\theta}_*)\\ \sqrt{n}S_1(\boldsymbol{\theta}_*)\end{bmatrix} + A^{22}(\boldsymbol{\theta}_*)\boldsymbol{\xi}.\end{aligned} \tag{A.16}$$

Under H_1, the asymptotic expected value of $S_2(\tilde{\boldsymbol{\theta}})$ is equal to $\frac{1}{\sqrt{n}}A^{22}(\boldsymbol{\theta}_*)\boldsymbol{\xi}$ and the asymptotic variance is the one in formulas (A.7) or (A.9).

Under H_1

$$\begin{aligned}S_2(\tilde{\boldsymbol{\theta}}_n) \sim N(\frac{1}{\sqrt{n}}A^{22}(\boldsymbol{\theta}_*)\boldsymbol{\xi}, B_{22}(\boldsymbol{\theta}_*) - A_{21}(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)B_{12}(\boldsymbol{\theta}_*) - B_{21}(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)A_{12}(\boldsymbol{\theta}_*) +\\ + A_{21}(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)B_{11}(\boldsymbol{\theta}_*)(\boldsymbol{\theta}_*)A_{11}^{-1}(\boldsymbol{\theta}_*)A_{12}(\boldsymbol{\theta}_*))\end{aligned} \tag{A.17}$$

and the $LM(S)$ statistic is asymptotically distributed as $\chi^2_r(\lambda)$ with non-centrality parameter given by the following formula (Bera et al., 2020)

$$\lambda = E(x)'V(x)^{-1}E(x). \tag{A.18}$$

Substituting the asymptotic mean and variance of $S_2(\tilde{\boldsymbol{\theta}}_n)$ in formula (A.18) we get

$$\lambda = \frac{1}{n}\boldsymbol{\xi}'A^{22'}(B_{22} - A_{21}A_{11}^{-1}B_{12} - B_{21}A_{11}^{-1}A_{12} + A_{21}A_{11}^{-1}B_{11}A_{11}^{-1}A_{12})^{-1}A^{22}\boldsymbol{\xi},$$

where all matrices are evaluated at $\boldsymbol{\theta}^*$. It can be easily verified that, in the absence of model misspecification, the non-centrality parameter of the LM(S) test reduces to formula (2.9).

A.3 Results of the simulation study under correct model specification

The same true parameter values, simulation conditions, unconstrained model and hypotheses introduced in section 2.5.1 are used under correct model specification to study the Type I error and power of the tests.

Table A.1 presents the Type I error rates of the LM(H), LM(CP), and LM(S) tests under correct model specification for scenarios **A**,**B** and **C**.

TABLE A.1: Type I error rates of the LM(H), LM(CP), and LM(S) tests under scenarios A, B and C, $p = 10, 20$, $n = 200, 500, 1000, 5000$

SC	p	n	LM(H)	LM(CP)	LM(S)
A	10	200	0.04	0.055	0.055
		500	0.04	0.06	0.05
		1000	0.07	0.08	0.065
		5000	0.075	0.07	0.07
	20	200	0.04	**0.125**	0.055
		500	0.055	0.07	0.055
		1000	0.04	0.05	0.04
B	10	200	**0.10**	**0.11**	0.045
		500	0.06	0.05	0.035
		1000	0.075	0.08	0.06
	20	200	0.05	**0.185**	0.055
		500	0.06	**0.09**	0.06
		1000	0.04	0.045	0.03
C	10	200	0.07	**0.085**	0.035
		500	0.045	0.07	0.045
		1000	0.08	**0.115**	**0.085**
		5000	0.07	0.08	0.08
	20	200	0.055	**0.155**	0.025
		500	0.06	0.075	0.06
		1000	0.045	0.06	0.05

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

Table A.2 presents the power of the LM(H), LM(CP), and LM(S) tests under correct model specification for scenarios **D**,**E** and **F**.

TABLE A.2: Power of the LM(H), LM(CP), and LM(S) tests under scenarios D, E and F, $p = 10, 20$, $n = 200, 500, 1000, 5000$

SC	p	n	LM(H)	LM(CP)	LM(S)
D	10	200	0.315	0.355	0.32
		500	0.675	0.705	0.675
		1000	0.925	0.94	0.93
		5000	1	1	1
	20	200	0.37	0.495	0.39
		500	0.755	0.795	0.76
		1000	0.97	0.97	0.97
E	10	200	0.39	0.525	0.355
		500	0.905	0.915	0.895
		1000	0.995	0.995	0.995
	20	200	0.755	0.885	0.76
		500	1	1	1
		1000	1	1	1
F	10	200	0.61	0.61	0.38
		500	0.935	0.92	0.9
		1000	0.995	0.99	0.99
	20	200	0.796	0.835	0.595
		500	0.985	0.985	0.98
		1000	1	1	1

A.4 Parameter bias under model misspecification

In this section some results on the mean bias for the ML estimates across replications under model misspecification are reported. In each scenario considered, the mean bias of each model parameter, generally named θ, is computed as:

$$Bias_{\hat{\theta}} = \frac{\sum_{l=1}^{R} |\hat{\theta}_l - \theta_0|}{R},$$

where θ_0 is the true parameter, $\hat{\theta}_l$ is the ML estimate of the parameter θ in the l-th replication and R is the number of replications, equal to 500.

Table A.3 presents the parameters bias under local dependence ($\sigma_u^2 = 2.25$ and $LD = 50\%$) for scenario **C**.

TABLE A.3: Parameters bias under local dependence ($\sigma_u^2 = 2.25$ and $LD = 50\%$), under scenario C, $p = 20$, $n = 1000$

θ	$Bias_{\hat{\theta}}$									
$\alpha_{0j}(j = 1, ..., 20)$	0.15	0.13	0.14	0.14	0.12	0.11	0.15	0.14	0.09	0.13
	0.27	0.17	0.37	0.41	0.19	0.34	0.33	0.18	0.34	0.35
$\alpha_{1j}(j = 1, ..., 20)$	0.57	0.17	0.46	0.45	0.24	0.13	0.46	0.32	0.10	0.42
	0.48	0.50	0.48	0.49	0.48	0.47	0.48	0.49	0.48	0.48
β	0.14									

Table A.4 presents the parameters bias under misspecification of the latent variable distribution ($\epsilon \sim SN(1)$, $\epsilon \sim SN(3)$) for scenario **A**.

TABLE A.4: Parameters bias under misspecification of the latent variable distribution ($\epsilon \sim SN(1)$, $\epsilon \sim SN(3)$), under scenario A, $p = 20$, $n = 1000$

$\epsilon \sim$	θ	$Bias_{\hat{\theta}}$									
SN(1)	$\alpha_{0j}(j = 1, ..., 20)$	1.19	0.61	1.06	1.05	0.74	0.49	1.04	0.87	0.35	1.04
		0.60	0.94	0.38	0.26	0.85	0.45	0.50	0.88	0.44	0.40
	$\alpha_{1j}(j = 1, ..., 20)$	0.43	0.22	0.39	0.38	0.27	0.18	0.37	0.32	0.14	0.37
		0.22	0.33	0.14	0.11	0.32	0.17	0.18	0.32	0.16	0.14
	β	0.23									
SN(3)	$\alpha_{0j}(j = 1, ..., 20)$	1.55	0.79	1.37	1.35	0.96	0.63	1.34	1.13	0.46	1.34
		0.78	1.22	0.49	0.34	1.10	0.59	0.66	1.14	0.57	0.52
	$\alpha_{1j}(j = 1, ..., 20)$	0.98	0.48	0.89	0.86	0.60	0.38	0.85	0.71	0.28	0.84
		0.48	0.75	0.29	0.20	0.70	0.36	0.40	0.72	0.35	0.32
	β	0.75									

Appendix B

Appendix

B.1 The mean and the variance of SNP latent variable

To compute the final estimator $\hat{\boldsymbol{\alpha}}_0$ in formula (3.9) and $\hat{\boldsymbol{\alpha}}_1$ in (3.10), it is necessary to compute $\tilde{E}(Z)$ and $\tilde{V}(Z)$ for the latent variable with density in (3.3). These quantities can be derived analytically.

We consider $L = 2$. After the optimization process, $P_L(z) = a_0 + a_1 z + a_2 z^2$ and $h(z|\boldsymbol{\varphi}^*) = P_L^2(z)\phi(z)$, where $a_0 = sin\varphi_1^* - \frac{1}{\sqrt{2}}cos\varphi_1^* cos\varphi_2^*$, $a_1 = cos\varphi_1^* sin\varphi_2^*$, $a_2 = \frac{1}{\sqrt{2}}cos\varphi_1^* cos\varphi_2^*$.

From Zhang and Davidian (2001)

$$\tilde{E}(Z) = a' M^* a \tag{B.1}$$

where the element in the i-th row and j-th column of M^* is $E(z^{i+j-1})$ and $z \sim N(0,1)$. The matrix M^* includes the moment of a standard normal distribution.

When $L = 2$

$$M^* = E\begin{pmatrix} z & z^2 & z^3 \\ z^2 & z^3 & z^4 \\ z^3 & z^4 & z^5 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 3 \\ 0 & 3 & 0 \end{pmatrix} \tag{B.2}$$

and

$$\begin{aligned}\tilde{E}(Z) &= \begin{pmatrix} sin\varphi_1^* - \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* & cos\varphi_1^*sin\varphi_2^* & \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 3 \\ 0 & 3 & 0 \end{pmatrix} \begin{pmatrix} sin\varphi_1 - \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \\ cos\varphi_1^*sin\varphi_2^* \\ \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \end{pmatrix} = \\ &= \begin{pmatrix} cos\varphi_1^*sin\varphi_2^* & sin\varphi_1^* + \frac{2}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* & 3cos\varphi_1^*sin\varphi_2^* \end{pmatrix} \begin{pmatrix} sin\varphi_1^* - \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \\ cos\varphi_1^*sin\varphi_2^* \\ \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \end{pmatrix} = \\ &= 2sin\varphi_1^*cos\varphi_1^*sin\varphi_2^* + \frac{4}{\sqrt{2}}cos\varphi_1^{*2}cos\varphi_2^*sin\varphi_2^* \end{aligned} \tag{B.3}$$

To compute $\tilde{V}(Z)$ we need also $\tilde{E}(Z^2)$. It can be computed as as $a'M^{**}a$, where the element in the i-th row and j-th column of M^{**} is $E(z^{i+j})$, and $z \sim N(0,1)$ (Zhang and Davidian, 2001). When $L = 2$

$$M^{**} = E\begin{pmatrix} z^2 & z^3 & z^4 \\ z^3 & z^4 & z^5 \\ z^4 & z^5 & z^6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 3 & 0 \\ 3 & 0 & 15 \end{pmatrix} \tag{B.4}$$

and

$$\begin{aligned}\tilde{E}(Z^2) &= \begin{pmatrix} sin\varphi_1^* - \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* & cos\varphi_1^*sin\varphi_2^* & \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 \\ 0 & 3 & 0 \\ 3 & 0 & 15 \end{pmatrix} \begin{pmatrix} sin\varphi_1^* - \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \\ cos\varphi_1^*sin\varphi_2^* \\ \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \end{pmatrix} = \\ &= \begin{pmatrix} sin\varphi_1^* + \frac{2}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* & 3cos\varphi_1^*sin\varphi_2^* & 3sin\varphi_1^* + \frac{12}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \end{pmatrix} \begin{pmatrix} sin\varphi_1^* - \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \\ cos\varphi_1^*sin\varphi_2^* \\ \frac{1}{\sqrt{2}}cos\varphi_1^*cos\varphi_2^* \end{pmatrix} = \\ &= sin\varphi_1^{*2} + \frac{4}{\sqrt{2}}cos\varphi_1^*sin\varphi_1^*cos\varphi_2^* + 3cos\varphi_1^{*2}sin\varphi_2^{*2} + 5cos\varphi_1^{*2}cos\varphi_2^{*2} \end{aligned} \tag{B.5}$$

The variance of the latent variable with a SNP density is computed as $\tilde{V}(Z) = \tilde{E}(Z^2) - \tilde{E}(Z)^2$, and we obtain

$$\begin{aligned}\tilde{V}(Z) = sin\varphi_1^{*2} + \frac{4}{\sqrt{2}}cos\varphi_1^*sin\varphi_1^*cos\varphi_2^* + 3cos\varphi_1^{*2}sin\varphi_2^{*2} + 5cos\varphi_1^{*2}cos\varphi_2^{*2} - \\ - (2sin\varphi_1^*cos\varphi_1^*sin\varphi_2^* + \frac{4}{\sqrt{2}}cos\varphi_1^{*2}cos\varphi_2^*sin\varphi_2^*)^2 \end{aligned} \tag{B.6}$$

To get the mean and the variance of the latent variable for the SNP_1 model, we should set $\varphi_2^* = \frac{\pi}{2}$ in equations (B.3) and (B.6).

B.2 Gradient computation: The SNP-IRT model

The function "nlminb" in R uses the analytically computed gradient. Following Irincheeva (2011), we can rewrite the likelihood in (3.4) as

$$l(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\int_R \left\{ \prod_{j=1}^{p} g_j \right\} P_L^2(z_i)\phi(z_i)dz_i \right], \tag{B.7}$$

where

$$g_j = \frac{\exp(y_{ij}\alpha_{0j} + y_{ij}\alpha_{1j}z_i)}{1 + \exp(\alpha_{0j} + \alpha_{1j}z_i)} \tag{B.8}$$

The derivative of the α_{0s} and α_{1s} parameter ($s = 1, .., p$) can be obtained with the following formula (Irincheeva et al., 2012):

$$\frac{\partial l(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_s} = \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1, j\neq s}^{p} g_j\} (\partial g_s / \partial \theta_s) P_L^2(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i}, \tag{B.9}$$

where $\theta_s = \alpha_{0s}$ or α_{1s}. The gradient with respect to these two parameters is obtained substituting the following two partial derivatives in (B.9):

$$\frac{\partial g_s}{\partial \alpha_{0s}} = y_{is}g_s - \frac{\exp(\alpha_{0s} + \alpha_{1s}z_i)}{1 + \exp(\alpha_{0s} + \alpha_{1s}z_i)} g_s \tag{B.10}$$

$$\frac{\partial g_s}{\partial \alpha_{1s}} = y_{is}g_s z_i - \frac{\exp(\alpha_{0s} + \alpha_{1s}z_i)}{1 + \exp(\alpha_{0s} + \alpha_{1s}z_i)} g_s z_i \tag{B.11}$$

Hence, the gradient of the α_{0s} parameter is:

$$\frac{\partial l(\mathbf{y}, \boldsymbol{\theta})}{\partial \alpha_{0s}} = \sum_{i=1}^{n} y_{is} - \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} \tag{B.12}$$

The gradient of the α_{1s} parameter is:

$$\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \alpha_{1s}} = \sum_{i=1}^{n} y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} - \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} \tag{B.13}$$

Let's consider $L = 2$. The analytical gradients for the $\boldsymbol{\varphi}$ parameter can be obtained with the following formula (Irincheeva et al., 2012):

$$\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \varphi} = \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1}^{p} g_j\} (2\partial P_L(z_i)/\partial \varphi) P_L(z_i) \phi(z_i) dz}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i}, \tag{B.14}$$

where $\varphi = \varphi_1$ or φ_2. The gradient with respect to these two parameters is obtained substituting the following two partial derivatives in (B.14):

$$\frac{\partial P_L(z_i)}{\partial \varphi_1} = cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z_i sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2} \tag{B.15}$$

$$\frac{\partial P_L(z_i)}{\partial \varphi_2} = cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2} \tag{B.16}$$

Hence, we obtain

$$\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \varphi_1} = \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z_i sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z_i) \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} \tag{B.17}$$

$$\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \varphi_2} = \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z^2 cos\varphi_1 sin\varphi_2/\sqrt{2}) P_L(z_i) \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} \tag{B.18}$$

We can obtain the gradient for $L = 1$ setting $\varphi_2 = \pi/2$ in all the formulas and not computing $\frac{\partial l}{\partial \varphi_2}$.

B.3 Hessian computation: The SNP-IRT model

B.3.1 Elements on the main diagonal of the Hessian matrix

The element on the main diagonal of the Hessian matrix corresponding to the parameters α_{0s} $(s = 1, .., p)$ is:

$$\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \alpha_{0s}^2} = \frac{\partial}{\partial \alpha_{0s}}\left(\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \alpha_{0s}}\right) = \frac{\partial}{\partial \alpha_{0s}}\left(\sum_{i=1}^{n} y_{is} - \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1}^{p} g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i}\right) \tag{B.19}$$

Through algebraic operations we obtain:

$$\begin{aligned}\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \alpha_{0s}^2} =& \sum_{i=1}^{n}\left(\frac{\int_R \{\prod_{j=1}^{p} g_j\}\frac{\exp(2\alpha_{0s}+2\alpha_{1s}z_i)}{(1+\exp(\alpha_{0s}+\alpha_{1s}z_i))^2} P_L^2(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} - \right. \\ & - \frac{\int_R \{\prod_{j=1}^{p} g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{(1+\exp(\alpha_{0s}+\alpha_{1s}z_i))^2} P_L^2(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} - \\ & \left. - \frac{\left(\int_R \{\prod_{j=1}^{p} g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i)\phi(z_i)dz_i\right)^2}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i)^2}\right)\end{aligned} \tag{B.20}$$

The element on the main diagonal of the Hessian matrix correspond to the parameters α_{1s} $(s = 1, .., p)$ is:

$$\begin{aligned}\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \alpha_{1s}^2} = \frac{\partial(\partial l(\mathbf{y},\boldsymbol{\theta})/\partial \alpha_{1s})}{\partial \alpha_{1s}} = \frac{\partial}{\partial \alpha_{1s}}\left(\sum_{i=1}^{n}\left(y_{is}\frac{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)z_i\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} - \right.\right. \\ \left.\left. - \frac{\int_R \{\prod_{j=1}^{p} g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i)z_i\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i}\right)\right.\end{aligned} \tag{B.21}$$

Through algebraic operations we obtain:

$$
\begin{aligned}
\frac{\partial l(\mathbf{y},\boldsymbol{\theta})}{\partial \alpha_{1s}^2} = \sum_{i=1}^{n} \Bigg(& y_{is}^2 \frac{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} + \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(2\alpha_{0s}+2\alpha_{1s}z)}{(1+\exp(\alpha_{0s}+\alpha_{1s}z_i))^2} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} - \\
& - 2y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} - \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{(1+\exp(\alpha_{0s}+\alpha_{1s}z_i))^2} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z) \phi(z) dz} - \\
& - y_{is}^2 \frac{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i)^2}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} - \frac{(\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i)^2}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz)^2} + \\
& + 2y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i \int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} \Bigg)
\end{aligned}
\tag{B.22}
$$

The last elements on the main diagonal of the Hessian matrix are the second derivatives of $l(\mathbf{y},\boldsymbol{\theta})$ with respect to the parameters φ_1, φ_2.

To obtain these elements we use the formula of the derivative of a quotient

$$
y' = \frac{\partial}{\partial x}\left[\frac{f(x)}{g(x)}\right] = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} = \frac{f'(x)}{g(x)} - \frac{f(x)g'(x)}{g(x)^2} \tag{B.23}
$$

where in our case

$$
\frac{f'(x)}{g(x)} = \sum_{i=1}^{n} \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2[(\partial(\partial P_L(z_i)/\partial\varphi)/\partial\varphi) P_L(z_i) + (\partial P_L(z_i)/\partial\varphi)^2] \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} \tag{B.24}
$$

and

$$
\frac{f(x)g'(x)}{g(x)^2} = \sum_{i=1}^{n} \frac{(\int_R \{\prod_{j=1}^{p} g_j\} (2\partial P_L(z_i)/\partial\varphi) P_L(z_i) \phi(z_i) dz_i)^2}{\left(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i \right)^2} \tag{B.25}
$$

We need

$$
\frac{\partial}{\partial\varphi_1}\left(\frac{\partial P_L(z_i)}{\partial\varphi_1}\right) = -sin\varphi_1 + cos\varphi_1 cos\varphi_2/\sqrt{2} - z_i cos\varphi_1 sin\varphi_2 - z_i^2 cos\varphi_1 cos\varphi_2/\sqrt{2} \tag{B.26}
$$

$$
\frac{\partial}{\partial\varphi_2}\left(\frac{\partial P_L(z_i)}{\partial\varphi_2}\right) = cos\varphi_1 cos\varphi_2/\sqrt{2} - z_i cos\varphi_1 sin\varphi_2 - z_i^2 cos\varphi_1 cos\varphi_2/\sqrt{2} \tag{B.27}
$$

Finally we obtain

$$\frac{\partial l}{\partial \varphi_1^2} = \sum_{i=1}^{n} \Bigg(\frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z_i sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2})^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} + \\ + \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(-sin\varphi_1 + cos\varphi_1 cos\varphi_2/\sqrt{2} - z_i cos\varphi_1 sin\varphi_2 - z_i^2 cos\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z)\phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} - \\ - \frac{(\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z_i sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z_i)\phi(z_i) dz_i)^2}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i)^2} \Bigg) \tag{B.28}$$

and

$$\frac{\partial l}{\partial \varphi_2^2} = \sum_{i=1}^{n} \Bigg(\frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2})^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} + \\ + \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 cos\varphi_2/\sqrt{2} - z_i cos\varphi_1 sin\varphi_2 - z_i^2 cos\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z)\phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} - \\ - \frac{(\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2}) P_L(z_i)\phi(z_i) dz_i)^2}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i)^2} \Bigg) \tag{B.29}$$

B.3.2 Elements outside the main diagonal of the Hessian matrix

For all this section $s, k = 1, ...p$ and $s \neq k$.

$$\frac{\partial l}{\partial \alpha_{0k}} \left(\frac{\partial l}{\partial \alpha_{0s}} \right) = \sum_{i=1}^{n} \Bigg(\frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}sz_i)} \frac{\exp(\alpha_{0k}+\alpha_{1k}z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k}z_i)} P_L^2(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} - \\ - \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i)\phi(z_i)dz_i \int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0k}+\alpha_{1k}z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k}z_i)} P_L^2(z_i)\phi(z_i)dz_i}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i)^2} \Bigg) \tag{B.30}$$

$$
\begin{aligned}
\frac{\partial l}{\partial \alpha_{1k}}\left(\frac{\partial l}{\partial \alpha_{1s}}\right) = \sum_{i=1}^{n} \Bigg(& y_{ik} y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} - y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0k}+\alpha_{1k} z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k} z_i)} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} - \\
& - y_{ik} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} + \\
& + \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} \frac{\exp(\alpha_{0k}+\alpha_{1k} z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k} z_i)} P_L^2(z_i) z_i^2 \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} - \\
& - y_{ik} y_{is} \frac{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i)^2}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} + \\
& + y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0k}+\alpha_{1k} z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k} z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i \int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} + \\
& + y_{ik} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i \int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} - \\
& - \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i \int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0k}+\alpha_{1k} z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k} z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} \Bigg)
\end{aligned}
\tag{B.31}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \alpha_{1k}}\left(\frac{\partial l}{\partial \alpha_{0s}}\right) = \sum_{i=1}^{n} \Bigg(& - y_{ik} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} + \\
& + \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} \frac{\exp(\alpha_{0k}+\alpha_{1k} z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k} z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i} + \\
& + y_{ik} \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} P_L^2(z_i) \phi(z_i) dz_i \int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} - \\
& - \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s} z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s} z_i)} P_L^2(z_i) \phi(z_i) dz_i \int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0k}+\alpha_{1k} z_i)}{1+\exp(\alpha_{0k}+\alpha_{1k} z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i}{(\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) \phi(z_i) dz_i)^2} \Bigg)
\end{aligned}
\tag{B.32}
$$

$$\begin{aligned}\frac{\partial l}{\partial\alpha_{1s}}\left(\frac{\partial l}{\partial\alpha_{0s}}\right)=\sum_{i=1}^{n}\Bigg(&-y_{is}\frac{\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)}P_L^2(z_i)z_i\phi(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}+\\
&+\frac{\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp(2\alpha_{0s}+2\alpha_{1s}z_i)}{(1+\exp(\alpha_{0s}+\alpha_{1s}z_i))^2}P_L^2(z_i)z_i\phi(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}-\\
&-\frac{\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{(1+\exp(\alpha_{0s}+\alpha_{1s}z_i))^2}P_L^2(z_i)z_i\phi(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}-\\
&+y_{is}\frac{\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)}P_L^2(z_i)\phi(z_i)dz_i\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)z_i\phi(z_i)dz_i}{(\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i)^2}-\\
&-\frac{\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)}P_L^2(z_i)\phi(z_i)dz_i\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)}P_L^2(z_i)z_i\phi(z_i)dz_i}{(\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i)^2}\Bigg)\end{aligned}\tag{B.33}$$

$$\begin{aligned}\frac{\partial l}{\partial\varphi_1}\left(\frac{\partial l}{\partial\varphi_2}\right)=\sum_{i=1}^{n}\Bigg(&\frac{\int_R\{\prod_{j=1}^{p}g_j\}2(cos\varphi_1sin\varphi_2/\sqrt{2}+z_icos\varphi_1cos\varphi_2-z_i^2cos\varphi_1sin\varphi_2/\sqrt{2})}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}\cdot\\
&\cdot(cos\varphi_1+sin\varphi_1cos\varphi_2/\sqrt{2}-z_isin\varphi_1sin\varphi_2-z_i^2sin\varphi_1cos\varphi_2/\sqrt{2})\phi(z_i)dz_i+\\
&+\frac{\int_R\{\prod_{j=1}^{p}g_j\}(-sin\varphi_1sin\varphi_2/\sqrt{2}-z_isin\varphi_1cos\varphi_2+z_i^2sin\varphi_1sin\varphi_2/\sqrt{2})2P_L(z_i)\phi(z_i)dz_i}{(\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}-\\
&-\frac{\int_R\{\prod_{j=1}^{p}g_j\}2(cos\varphi_1sin\varphi_2/\sqrt{2}+z_icos\varphi_1cos\varphi_2-z_i^2cos\varphi_1sin\varphi_2/\sqrt{2})P_L(z_i)\phi(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}\cdot\\
&\cdot\frac{\int_R\{\prod_{j=1}^{p}g_j\}2(cos\varphi_1+sin\varphi_1cos\varphi_2/\sqrt{2}-z_isin\varphi_1sin\varphi_2-z_i^2sin\varphi_1cos\varphi_2/\sqrt{2})P_L(z_i)\phi(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}\Bigg)\end{aligned}\tag{B.34}$$

$$\begin{aligned}\frac{\partial l}{\partial\varphi_1}\left(\frac{\partial l}{\partial\alpha_{0s}}\right)=\sum_{i=1}^{n}\Bigg(&\\
&-\frac{\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)}2(cos\varphi_1+sin\varphi_1cos\varphi_2/\sqrt{2}-z_isin\varphi_1sin\varphi_2-z_i^2sin\varphi_1cos\varphi_2/\sqrt{2})P_L(z_i)\phi(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}+\\
&+\frac{\int_R\{\prod_{j=1}^{p}g_j\}\frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)}P_L^2(z_i)\phi(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j\}P_L^2(z_i)\phi(z_i)dz_i}\cdot\\
&\cdot\frac{\int_R\{\prod_{j=1}^{p}g_j\}2(cos\varphi_1+sin\varphi_1cos\varphi_2/\sqrt{2}-zsin\varphi_1sin\varphi_2-z_i^2sin\varphi_1cos\varphi_2/\sqrt{2})\phi(z_i)P_L(z_i)dz_i}{\int_R\{\prod_{j=1}^{p}g_j)\}P_L(z_i)^2\phi(z_i)dz_i}\Bigg)\end{aligned}\tag{B.35}$$

$$\frac{\partial l}{\partial \varphi_2}\left(\frac{\partial l}{\partial \alpha_{0s}}\right) =$$

$$= \sum_{i=1}^{n} \Bigg(- \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2}) P_L(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} +$$

$$+ \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} \cdot$$

$$\cdot \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2})\phi(z_i) P_L(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L(z_i)^2\phi(z_i)dz_i} \Bigg) \tag{B.36}$$

$$\frac{\partial l}{\partial \varphi_1}\left(\frac{\partial l}{\partial \alpha_{1s}}\right) =$$

$$\sum_{i=1}^{n} \Bigg(y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z_i sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z_i) z_i \phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} -$$

$$- \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} 2(cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z_i sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z_i) z_i \phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} -$$

$$- y_{is} \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z_i sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} \cdot$$

$$\cdot \frac{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} +$$

$$+ \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 + sin\varphi_1 cos\varphi_2/\sqrt{2} - z sin\varphi_1 sin\varphi_2 - z_i^2 sin\varphi_1 cos\varphi_2/\sqrt{2}) P_L(z_i)\phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} \cdot$$

$$\cdot \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i) z_i \phi(z_i)dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i)dz_i} \Bigg) \tag{B.37}$$

$$
\begin{aligned}
&\frac{\partial l}{\partial \varphi_2}\left(\frac{\partial l}{\partial \alpha_{1s}}\right) = \\
&\sum_{i=1}^{n}\Bigg(y_{is}\frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2}) P_L(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i) dz_i} - \\
&- \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2}) P_L(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i) dz_i} - \\
&- y_{is}\frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2}) P_L(z_i) \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i) dz_i} \cdot \\
&\cdot \frac{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i) dz_i} + \\
&+ \frac{\int_R \{\prod_{j=1}^{p} g_j\} 2(cos\varphi_1 sin\varphi_2/\sqrt{2} + z_i cos\varphi_1 cos\varphi_2 - z_i^2 cos\varphi_1 sin\varphi_2/\sqrt{2}) P_L(z_i) \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i) dz_i} \cdot \\
&\cdot \frac{\int_R \{\prod_{j=1}^{p} g_j\} \frac{\exp(\alpha_{0s}+\alpha_{1s}z_i)}{1+\exp(\alpha_{0s}+\alpha_{1s}z_i)} P_L^2(z_i) z_i \phi(z_i) dz_i}{\int_R \{\prod_{j=1}^{p} g_j\} P_L^2(z_i)\phi(z_i) dz_i}\Bigg)
\end{aligned}
\tag{B.38}
$$

We can obtain the Hessian matrix for $L = 1$ setting $\varphi_2 = \pi/2$ in all the formulas and not computing the derivatives that involve the parameter φ_2.

Bibliography

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). Wiley.

Bartolucci, F., Bacci, S., & Pigini, C. (2017). Misspecification test for random effects in generalized linear finite-mixture models for clustered binary and ordered data. *Econometrics and Statistics*, *3*, 112–131.

Bera, A. K., Bilias, Y., Yoon, M. J., Taşpınar, S., & Doğan, O. (2020). Adjustments of Rao's score test for distributional and local parametric misspecifications. *Journal of Econometric Methods*, *9*(1), 20170022.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

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

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. *Essential statistical inference: Theory and methods* (pp. 339–359). Springer.

Cagnone, S., & Viroli, C. (2012). A factor mixture analysis model for multivariate binary data. *Statistical Modelling*, *12*(3), 257–277.

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

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Davidian, M., & Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, *80*(3), 475–488.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*(4), 295–311.

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. *Annals of Mathematical Statistics*, *28*(1), 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., pp. 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. W. (2005). Bayesian modification indices for IRT models. *Statistica Neerlandica*, *59*(1), 95–106.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*(2), 363–390.

Gallant, A. R., & Tauchen, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica*, *57*(5), 1091–1120.

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

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

Glas, C. A. W., & 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*, *54*(2), 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. (2021). The asymptotic power of the Lagrange multiplier tests for misspecified IRT models. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative Psychology: The 85th annual meeting of the Psychometric Society, virtual* (pp. 275–284). Springer.

Guastadisegni, L., Cagnone, S., Moustaki, I., & Vasdekis, V. (2022). Use of the Lagrange multiplier test for assessing measurement invariance under model misspecification. *Educational and Psychological Measurement*, *82*(2), 254–280.

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.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 190–195.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251–1271.

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

Irincheeva, I., Cantoni, E., & Genton, M. G. (2012). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, *39*(4), 663–680.

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.

Knott, M., & Tzamourani, P. (2007). Bootstrapping the estimated latent distribution of the two-parameter latent trait model. *British Journal of Mathematical and Statistical Psychology*, *60*(1), 175–191.

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*(370), 336–344.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge.

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

Ma, Y., & Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 475–495.

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

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

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.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101.

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 S. H. Irvine & W. J. Berry (Eds.), *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.

Monroe, S. L. (2014). *Multidimensional item factor analysis with semi-nonparametric latent densities* (Doctoral dissertation). UCLA.

Montanari, A., & Viroli, C. (2010). A skew-normal factor model for the analysis of student satisfaction towards university courses. *Journal of Applied Statistics*, *37*(3), 473–487.

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.

Ranger, J., & Much, S. (2020). Analyzing the fit of IRT models with the Hausman test. *Frontiers in Psychology*, *11*.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*(1), 50–57.

Rao, J., Scott, A. J., & Skinner, C. J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, *8*, 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*, *61*(2), 331–360.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph No. 17*.

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

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

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.

Tchetgen, E. J., & Coull, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika*, *93*(4), 1003–1010.

van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika*, *81*(3), 650–673.

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

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*(3-4), 350–362.

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

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, *11*(3), 253–270.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102–117.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*(5), 737–757.

Yuan, K.-H., & Bentler, P. M. (2010). Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 273–291.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, *57*(3), 795–802.