Received: 1 August 2023 | Accepted: 1 December 2024

DOI: 10.1111/bmsp.12379

the british psychological society
promoting excellence in psychology

ARTICLE

The generalized Hausman test for detecting non-normality in the latent variable distribution of the two-parameter IRT model

Lucia Guastadisegni[1] | **Silvia Cagnone**[1] | **Irini Moustaki**[2] | **Vassilis Vasdekis**[3]

[1]University of Bologna, Bologna, Italy

[2]London School of Economics and Political Science, London, UK

[3]Athens University of Economics and Business, Athens, Greece

Correspondence
Silvia Cagnone, Department of Statistical Sciences "Paolo Fortunati", Via Belle Arti 41, Bologna, Italy.
Email: silvia.cagnone@unibo.it

Funding information
European Union-NextGenerationEU, GRINS-Growing Resilient, INclusive and Sustainable project, Grant/Award Number: PNRR-M4C2-I1.3-PE00000018-CUPJ33C22002910001

Abstract

This paper introduces the generalized Hausman test as a novel method for detecting the non-normality of the latent variable distribution of the unidimensional latent trait model for binary data. The test utilizes the pairwise maximum likelihood estimator for the parameters of the latent trait model, which assumes normality of the latent variable, and the maximum likelihood estimator obtained under a semi-non-parametric framework, allowing for a more flexible distribution of the latent variable. The performance of the generalized Hausman test is evaluated through a simulation study and compared with other test statistics available in the literature for testing latent variable distribution fit and an overall goodness-of-fit test statistic. Additionally, three information criteria are used to select the best-fitted model. The simulation results show that the generalized Hausman test outperforms the other tests under most conditions. However, the results obtained from the information criteria are somewhat contradictory under certain conditions, suggesting a need for further investigation and interpretation. The proposed test statistics are used in three datasets.

KEYWORDS

correlated binary data, misspecification test, semi-non-parametric-IRT model

1 | INTRODUCTION

This paper proposes a generalized Hausman-type test for detecting deviation of the latent variable distribution from normality in the unidimensional latent trait model, also known as the two-parameter model (2PL) in item response theory (IRT) (see e.g. van der Linden & Hambleton, 2013). The paper focuses on a mixture of normals and skew-normal latent distributions.

Latent variable or trait models are used to study theoretical constructs, such as abilities, attitudes, quality of life or business confidence, which cannot be directly observed and measured. Latent variables are here continuous variables measured through correlated categorical indicators known as manifest variables or items (Bartholomew et al., 2011). This paper considers the two-parameter latent trait model for binary manifest variables (2PL).

One of the standard assumptions of IRT models is that the latent variable(s) follows a normal distribution. However, assuming normality of the latent variable(s) when the actual distribution has a different shape can lead to biased parameter estimates, especially with binary outcomes (Ma & Genton, 2010). Furthermore, assuming an incorrect distribution of the latent variable can lead to erroneous conclusions when conducting hypothesis testing (Guastadisegni et al., 2022).

The need to accommodate more flexible latent variable distributions has been recognized in the literature. Montanari and Viroli (2010) introduced a skew-normal latent variable in the classical factor analysis model, while Cagnone and Viroli (2012) presented a latent trait model where the factors are distributed as a finite mixture of multivariate Gaussians. Within the generalized linear latent variable model (GLLVM) framework (see e.g. Bartholomew et al., 2011; Skrondal & Rabe-Hesketh, 2004), Ma and Genton (2010) proposed a semi-parametric method, consistent for various types of manifest variables under different distributions of the latent variables, and Irincheeva et al. (2012) considered the semi-non-parametric (SNP) approach, introduced by Gallant and Nychka (1987). This approach allows for more flexible smooth densities of the latent variables. The SNP method has also been used in the unidimensional 2PL model by Woods and Lin (2009) and in the multidimensional 2PL model by Monroe (2014). For the 2PL model, Bock and Aitkin (1981) modelled the density of the latent variable using an empirical histogram instead of assuming a specific parametric form. They estimated the density together with the item parameters within an expectation-maximization (E-M) algorithm. The density of the latent variable was estimated directly at each quadrature point, rather than being derived from a predefined normal probability distribution. Woods and Thissen (2006) proposed a non-parametric estimation of the latent variable while maintaining the logistic item response function (IRF). They modelled the latent variable through a Ramsey curve, which is a spline-based density (Ramsay, 2000).

In the majority of cases, information criteria, such as the Akaike information criterion (AIC) (Akaike, 1974) or the Bayesian information criterion (BIC) (Schwarz, 1978), 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 & Lin, 2009; Irincheeva et al., 2012; Monroe, 2014). With continuous manifest variables, Ma and Genton (2010) perform the Kolmogorov–Smirnov test on the continuous responses' distribution to evaluate the latent variable's normality.

When dealing with categorical responses, Li and Cai (2018) suggested employing summed score likelihood-based statistics to test the fit of the latent variable distribution. They proposed two versions of the test: an unadjusted version, $\overline{X}^2$, and a moment-adjusted version (Satorra & Saris, 1985). Similarly, Monroe (2021) introduced an alternative test statistic, denoted here by $R(z)$, for detecting non-normality in the latent variable distribution across the range of z. It is based on generalized residuals (Haberman & Sinharay, 2013), which compare the specified latent variable distribution to the sample average of latent variable posterior distributions, asymptotically distributed as standard normal. The paper also proposed a version to test the hypothesis that the distribution of the latent variable as a whole is correctly specified in the model denoted here as R_B. A Bonferroni correction was applied to the critical value. The statistics proposed by Li and Cai (2018) and Monroe (2021) demonstrate high power when the latent variable is skewed or platykurtic and are insensitive to multidimensionality.

Hausman (1978) proposed a specification test to detect the failure of the orthogonality assumption in regression analysis. Due to its simplicity, the Hausman test can be applied in various contexts to detect different types of model misspecification. This test compares two different estimators that are consistent when the model is correctly specified, and one is also efficient. In the presence of model misspecification, only the inefficient estimator is consistent. The efficiency assumption simplifies the covariance matrix computation of the difference between the two estimators. However, this matrix can fail to be positive definite under model misspecification or in small sample sizes. Moreover, neither of the two estimators is considered fully efficient in some cases.

The generalized version of the Hausman (GH) test, proposed by White (1982), is a more flexible and robust alternative to the original Hausman test. Indeed, the generalized version allows both estimators to be inefficient and obtained from two different models. Moreover, the covariance matrix of the difference between the two estimators is robust and always positive definite.

In the IRT context, as far as we know, the 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 highlight that this test performs well regarding Type I error rates for large sample sizes and power under most conditions. 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 also be used to detect the possible correlation between random effects and cluster-specific covariates and detect endogeneity.

This work aims to extend the GH test to detect the non-normality of the latent variable distribution in a unidimensional latent trait model for binary data. The estimators resulting from two different models are considered to build the test. The first model is a 2PL unidimensional IRT model that assumes the normality of the latent variable. In contrast, the second model is the unidimensional SNP-IRT model, which assumes a more flexible distribution for the latent variable. The 2PL model is estimated using a pairwise maximum likelihood (PL) method (Katsikatsou et al., 2012), which is a composite likelihood method that uses information from bivariate-order margins (Lindsay, 1988; Varin, 2008). The SNP-IRT model is estimated using a maximum likelihood (ML) method. The choice of these estimators and models is motivated by the following reasons. First, both estimators are consistent when the latent variable is normally distributed. Moreover, the ML estimator for the SNP-IRT model is also consistent under different distribution assumptions of the latent variable (Gallant & Tauchen, 1989; Irincheeva et al., 2012). These conditions on the consistency of the parameter estimators are required to correctly apply the GH test (White, 1982). Second, the PL and ML estimators yield different variances for the estimated parameters. This implies that under the normality of the latent variable distribution, the covariance matrix of the difference between the two estimators involved in the computation of the GH test differs from zero. The non-zero covariance matrix avoids numerical instability in the calculation of the test.

The theoretical aspects of the models, estimators and matrices involved in the computation of the GH test are described in the following sections. Moreover, we carry out an extensive simulation study to evaluate the performance of the GH test in terms of both Type I error rates and empirical power. For the latter, we consider both a mixture of normals and skew-normal distributions for the latent variable, with varying degrees of departure from the normal distribution. Additionally, we evaluate the asymptotic behaviour of the test in terms of both Type I error rates and power, using a very large sample size.

Furthermore, the performance of the GH test is compared with three available test statistics. Namely, the M_2 is a statistic based on standardized univariate and bivariate residuals unaffected by sparseness (Maydeu-Olivares & Joe, 2005). The likelihood ratio (LR) test for nested models and the $\overline{X^2}$ statistic proposed by Li and Cai (2018). The latter is available in FlexMIRT (Cai, 2017). We also study the performance of three model selection information criteria. Three applications to real data are also presented.

This article is organized as follows. In Section 2, we describe the 2PL model and SNP-IRT models for binary data and the PL and ML estimation, respectively. Section 3 presents the GH test to detect the non-normality of the latent variable distribution. In Section 4, we review the M_2, the LR, the $\overline{X^2}$ and the R_B test statistics, and in Section 5, the information criteria. Section 6 presents a Monte Carlo simulation

study, and Section 7 presents the results from three real data analyses. Finally, in Section 8, some concluding remarks are presented and discussed.

2 | THE 2PL AND SNP-IRT MODEL FOR BINARY DATA

Consider n respondents answering p binary items. Let $y_{ij} \in \{0,1\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$ be a binary random variable recording individual i's response to item j. We assume that the items measure a unidimensional construct modelled by a latent variable, z. The density function of this variable is expressed as $h(z)$.

According to the two-parameter model (2PL), the probability of a positive response for y_{ij} is modelled using a logistic model (measurement model) given by

$$P(y_{ij} = 1 | z_i) = \pi_{ij}(z_i) = \frac{\exp(\alpha_{0j} + \alpha_{1j} z_i)}{1 + \exp(\alpha_{0j} + \alpha_{1j} z_i)}, \; j = 1, \ldots, p \tag{1}$$

where α_{0j} is the item intercept and α_{1j} the item slope (factor loading) and $h(z) = \phi(z)$, where $\phi(z)$ is the standard normal density.

An extension of the 2PL model is given by the SNP-IRT model that assumes the same response probability as (1) but a SNP parametrization of the latent variable as follows:

$$h(z_i) = P_L^2(z_i)\phi(z_i) \text{ and } P_L(z_i) = \sum_{0 \le l \le L} a_l z_i^l, \tag{2}$$

where $a_0, \ldots, a_L$ are the real coefficients of the polynomial $P_L(z_i)$ and L is the polynomial degree.

For $h(z)$ to be a density, the coefficients $a_0, \ldots, a_L$ of $P_L(z)$ should be chosen such that $\int h(z)dz = 1$.

We follow the same parameterization as Woods and Lin (2009) and Irincheeva et al. (2012) to obtain a unique representation of the polynomial coefficients in Equation (2). The details of this parameterization can be found in Irincheeva et al. (2012) and are reported in the Data S1: Section S1.1.

Here, we consider L = 0, 1 and 2. When $L = 0$, the distribution of the latent variable in (2) reduces to the standard normal one. When $L = 1$, $P_L(z) = a_0 + a_1 z$, where $a_0 = \sin\varphi_1$, $a_1 = \cos\varphi_1$, with $-\pi/2 < \varphi_1 \le \pi/2$. The SNP parametrization with $L = 1$ includes unimodal and bimodal distributions.

Figure 1 displays the SNP densities of z when $L = 1$, for different values of the φ_1 parameter.

When φ_1 is negative and close to -1, the distribution is slightly right-skewed, whereas when it is close to 1, it is slightly left-skewed. When the values of φ_1 are between -1 and 1, the distributions are bimodal. Even if not reported in the graph, when $\varphi_1 = \frac{\pi}{2}$, the SNP distribution reduces to the standard normal.

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 $a_2 = \frac{1}{\sqrt{2}}\cos\varphi_1\cos\varphi_2$, with $-\pi/2 < \varphi_l \le \pi/2$, $l = 1, 2$. The SNP parametrization with $L = 2$ is more flexible than $L = 1$ and encompasses unimodal, multimodal (including up to three modes) and skewed distributions. Figure 2 displays the SNP densities of z when $L = 2$, for different values of the φ_1 and φ_2 parameters.

When the value of φ_1 is negative, and the value of φ_2 is close to -1 or 1, the distributions are bimodal. When both parameters are close to 0, the distributions are trimodal. When the value of φ_1 is positive and >0.5, and the value of φ_2 is close to -1 or 1, the distributions are highly right-skewed and left-skewed, respectively. Even though it is not reported in Figure 2, when $\varphi_2 = \frac{\pi}{2}$, the SNP densities reduce to those observed when $L = 1$. Moreover, when both parameters are set to $\frac{\pi}{2}$, the SNP densities return to the standard normal.

We indicate with SNP_2 the 2PL model for $L = 2$, with SNP_1 the 2PL model for $L = 1$ and with SNP_0 the 2PL model for $L = 0$.

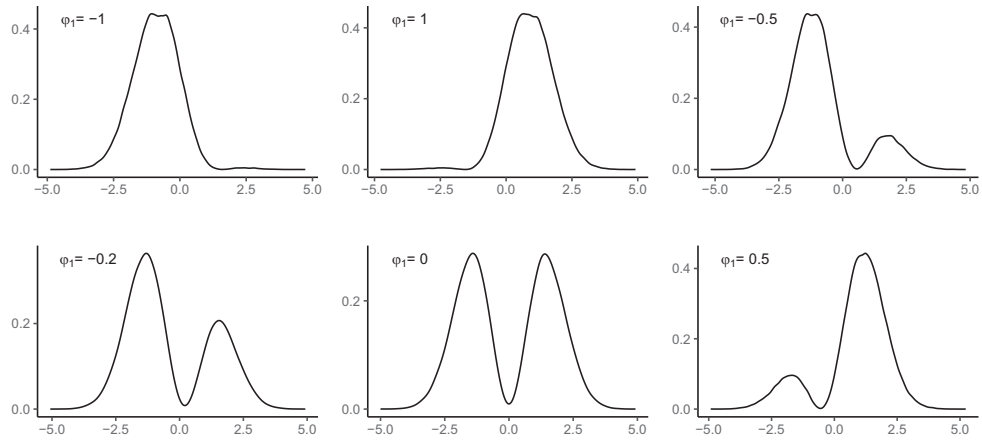

FIGURE 1 SNP densities of z when $L = 1$, for different values of the φ_1 parameter.

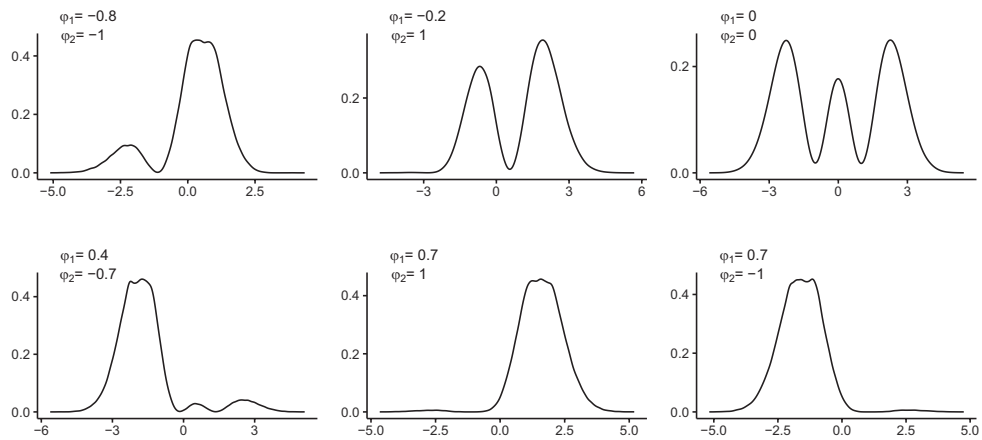

FIGURE 2 SNP densities of z when $L = 2$, for different values of the φ_1 and φ_2 parameters.

2.1 | PL estimation for the $\mathbf{SNP_0}$ model

To implement the GH test, the parameters of the SNP_0 model are estimated with the PL. The pairwise log-likelihood, based on the bivariate marginal densities $f(y_{ij}, y_{ik}, \boldsymbol{\theta})$, $j, k = 1, \ldots p$ and $k > j$, is

$$\begin{aligned} pl_{\mathrm{SNP}_0}(\mathbf{y}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k>j} \ln f(y_{ij}, y_{ik}, \boldsymbol{\theta}) = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k>j} \ln \int \left[\pi_{ij}(z_i)^{y_{ij}} (1 - \pi_{ij}(z_i))^{1-y_{ij}} \right] \left[\pi_{ik}(z_i)^{y_{ik}} (1 - \pi_{ik}(z_i))^{1-y_{ik}} \right] \phi(z_i) dz_i. \end{aligned} \tag{3}$$

$pl_{\mathrm{SNP}_0}(\mathbf{y}, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ includes the item intercepts and slopes. Under correct model specification, the maximum PL estimator $\tilde{\boldsymbol{\theta}}$ converges in probability to the true parameter vector $\boldsymbol{\theta}_0' = (\boldsymbol{\alpha}_{00}', \boldsymbol{\alpha}_{01}')$ and

$$\tilde{\boldsymbol{\theta}} \xrightarrow{p} N(\boldsymbol{\theta}_0, A^{-1}(\boldsymbol{\theta}_0) B(\boldsymbol{\theta}_0) A^{-1}(\boldsymbol{\theta}_0)), \tag{4}$$

where $\mathcal{A}(\boldsymbol{\theta}) = E_y\left[-\frac{\partial^2 pl_{\mathrm{SNP}_0}(\mathbf{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \right]$, $B = var\left[\frac{\partial pl_{\mathrm{SNP}_0}(\mathbf{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \right]$ and $\mathcal{A}(\boldsymbol{\theta}) \neq B(\boldsymbol{\theta})$ (Lindsay, 1988, Varin, 2008).

These matrices can be estimated by their observed versions evaluated at $\tilde{\boldsymbol{\theta}}$ as

$$\hat{\mathcal{A}}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \frac{\partial^2 pl_{\mathrm{SNP}_0}(\mathbf{y}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \tag{5}$$

and

$$\hat{B}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \frac{\partial pl_{\mathrm{SNP}_0}(\boldsymbol{y}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial pl_{\mathrm{SNP}_0}(\boldsymbol{y}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}. \tag{6}$$

2.2 | ML estimator for the SNP_L model

The parameter vector $\boldsymbol{\theta}^{(1)\prime} = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', \boldsymbol{\varphi}') = (\boldsymbol{\theta}', \boldsymbol{\varphi}')$ of the SNP_L model, where $L > 0$, is estimated using the ML method. The log-likelihood of the data is

$$\begin{aligned} l_{\mathrm{SNP}_L}(\mathbf{y},\boldsymbol{\theta}^{(1)}) &= \sum_{i=1}^{n} \ln f(\mathbf{y}_i,\boldsymbol{\theta}^{(1)}) = \\ &= \sum_{i=1}^{n} \ln \int \prod_{j=1}^{p} \pi_{ij}(z_i)^{y_{ij}}(1-\pi_{ij}(z_i))^{1-y_{ij}} P_L^2(z_i)\exp\Big(-\frac{1}{2}z_i'z_i\Big)dz_i. \end{aligned} \tag{7}$$

The integral in $l_{\mathrm{SNP}_L}(\mathbf{y},\boldsymbol{\theta}^{(1)})$ is approximated by using the Gauss–Hermite quadrature, as in Woods and Lin (2009). The degree of the polynomial L is fixed and is not estimated by ML. The log-likelihood function is maximized with respect to the unknown vector of parameter $\boldsymbol{\theta}^{(1)}$ as follows:

$$(\dot{\boldsymbol{\alpha}}_0', \dot{\boldsymbol{\alpha}}_1', \hat{\boldsymbol{\varphi}}') = argmax_{\boldsymbol{\theta}^{(1)}} l_{\mathrm{SNP}_L}(\mathbf{y},\boldsymbol{\theta}^{(1)}). \tag{8}$$

When the model is correctly specified, this is the full ML method. Under model misspecification, the method becomes a quasi-ML method (White, 1982). The two methods are computationally the same, but the estimators exhibit different theoretical properties, as will be clarified later in this section.

Regarding the identifiability of the model parameters with respect to the arbitrariness of the location and scale of the latent variable, we show in Section S1.4 of Data S1 that model parameters for the SNP_1 model are locally identifiable without having to fix the location and scale of the latent variable or the elements of the loading matrix to specified values. This is not the case for most of the latent variable models with a parametric distribution for the latent variable. In the same way, for larger SNP models, one can show that high-order moments of the observed variable contribute to the estimation of model parameters, thus providing at least locally identified model parameters. Irincheeva et al. (2012) also discusses the identifiability due to rotational indeterminacy.

After maximizing the log-likelihood in (7), the algorithm converges to $(\dot{\boldsymbol{\alpha}}_0', \dot{\boldsymbol{\alpha}}_1', \hat{\boldsymbol{\varphi}}')$ and the latent variable z has a density $h(z|\hat{\boldsymbol{\varphi}}, L)$ with an estimated mean $\tilde{E}(Z)$ and variance $\tilde{V}(Z)$ (see Section S1.2 of Data S1 for details). The estimators $\dot{\boldsymbol{\alpha}}_0$ and $\dot{\boldsymbol{\alpha}}_1$ are, in effect, calibrated with respect to the $h(z|\hat{\boldsymbol{\varphi}}, L)$ distribution. However, the estimates can be rescaled to compare with those of the 2PL with a standard normal latent variable distribution, as discussed in Section S1.3 of Data S1. The rescaled parameters are denoted as $\hat{\boldsymbol{\theta}}^{(1)\prime} = (\hat{\boldsymbol{\alpha}}_0', \hat{\boldsymbol{\alpha}}_1', \hat{\boldsymbol{\varphi}}')$.

The SNP densities can approximate various distributions, including a mixture of normals and skew distributions. However, as in almost all cases the SNP densities approximate the true latent variable distribution but do not exactly coincide with it, if the regularity conditions A2–A6 of White (1982) are satisfied, the obtained estimator is a Quasi-ML estimator

$$\widehat{\boldsymbol{\theta}}^{(1)} \xrightarrow{p} N\left(\boldsymbol{\theta}_0^{(1)}, \mathcal{A}^{-1}\left(\boldsymbol{\theta}_0^{(1)}\right) B\left(\boldsymbol{\theta}_0^{(1)}\right) \mathcal{A}^{-1}\left(\boldsymbol{\theta}_0^{(1)}\right)\right), \tag{9}$$

where $\boldsymbol{\theta}_0^{(1)\prime} = (\boldsymbol{\alpha}_{00}', \boldsymbol{\alpha}_{01}', \boldsymbol{\varphi}_*') = (\boldsymbol{\theta}_0', \boldsymbol{\varphi}_*')$. $\boldsymbol{\varphi}_*$ is the value of $\boldsymbol{\varphi}$ that minimizes the Kullback–Leibler information criterion (White, 1982; Gallant & Tauchen, 1989; Irincheeva et al., 2012). If the true latent variable follows exactly an SNP_L density, including the normal as a sub-case, that is, when $\varphi_l = \frac{\pi}{2}$, the vector $\boldsymbol{\varphi}_*$ coincides with the true parameter value $\boldsymbol{\varphi}_0$, and the quasi-ML method reduces to the full ML method. $\mathcal{A}(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ are the expected Hessian and cross-product matrices, respectively. Their observed versions can be computed with the Delta method (Cramér, 1946) and are defined similar to (5) and (6), where $pl_{\text{SNP}_0}(\mathbf{y}_i, \boldsymbol{\theta})$ is replaced by $l_{\text{SNP}_L}(\mathbf{y}_i, \boldsymbol{\theta}^{(1)})$.

3 | THE GENERALIZED HAUSMAN TEST

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

As in the previous sections, let us denote by $\boldsymbol{\theta}$ the sub-vector of $\boldsymbol{\theta}^{(1)\prime} = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', \boldsymbol{\varphi}')$ that includes the item intercepts $\boldsymbol{\alpha}_0$ and slopes $\boldsymbol{\alpha}_1$. $\boldsymbol{\theta}$ has dimension $2p \times 1$, where p is the number of items.

For the 2PL IRT model (SNP_0), consider the maximum PL estimator $\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}$.

Consider the ML estimator $\widehat{\boldsymbol{\theta}}^{(1)\prime}_{\text{SNP}_L} = (\widehat{\boldsymbol{\theta}}'_{\text{SNP}_L}, \widehat{\boldsymbol{\varphi}}')$ of a SNP-IRT model with $L > 0$, where the sub-vector of parameter $\widehat{\boldsymbol{\varphi}}$ has dimension $L \times 1$ and so $\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L}$ has dimension $(2p + L) \times 1$.

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

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{\text{SNP}_L} - \tilde{\boldsymbol{\theta}}_{\text{SNP}_0}\right) \xrightarrow{d} N\left(0, S\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0^{(1)}\right)\right). \tag{10}$$

An estimator of $S(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0{}^{(1)})$ is given by

$$\begin{aligned}\hat{S}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}, \widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L}) = {} & \hat{\mathcal{A}}^{\boldsymbol{\theta}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})^{-1} \hat{B}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L}) \hat{\mathcal{A}}^{\boldsymbol{\theta}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})^{-1\prime} + \hat{\mathcal{A}}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0})^{-1} \hat{B}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}) \hat{\mathcal{A}}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0})^{-1\prime} - \\ & - \hat{\mathcal{A}}^{\boldsymbol{\theta}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})^{-1} \hat{R}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}, \widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})' \hat{\mathcal{A}}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0})^{-1\prime} - \hat{\mathcal{A}}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0})^{-1} \hat{R}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}, \widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L}) \hat{\mathcal{A}}^{\boldsymbol{\theta}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})^{-1\prime},\end{aligned} \tag{11}$$

where the matrices $\hat{\mathcal{A}}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0})$ and $\hat{B}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0})$, defined in Equations (5) and (6), have dimension $2p \times 2p$ and are evaluated at $\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}$. $\hat{\mathcal{A}}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})$ and $\hat{B}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})$ are the observed Hessian and cross-product matrix of dimension $(2p + L) \times (2p + L)$ for the SNP_L model, evaluated at $\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L}$. The matrix $\hat{\mathcal{A}}^{\boldsymbol{\theta}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})^{-1}$ is obtained by deleting the last L rows from the matrix $\hat{\mathcal{A}}(\widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})^{-1}$ and has dimension $2p \times (2p + L)$. The matrix $\hat{R}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}, \widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})$ has dimension $2p \times (2p + L)$ and can be computed as

$$\hat{R}(\boldsymbol{\theta}_{\text{SNP}_0}, \boldsymbol{\theta}^{(1)}_{\text{SNP}_L}) = \sum_{i=1}^{n} \frac{\partial pl_{\text{SNP}_0}(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial l_{\text{SNP}_L}(\boldsymbol{y}_i, \boldsymbol{\theta}^{(1)})}{\partial \boldsymbol{\theta}^{(1)\prime}}, \tag{12}$$

where $pl_{\text{SNP}_0}(\boldsymbol{y}_i, \boldsymbol{\theta})$ is the pairwise log-likelihood for the individual i under the model SNP_0 and $l_{\text{SNP}_L}(\boldsymbol{y}_i, \boldsymbol{\theta}^{(1)})$ is the log-likelihood for the individual i under the model SNP_L. The matrix in (12) is evaluated at $(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}, \widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})$. We choose the maximum PL and the ML estimator for the two models to avoid that, under correct model specification, $\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}$ and $\widehat{\boldsymbol{\theta}}_{\text{SNP}_L}$ converge to the same covariance matrix, producing a $\hat{S}(\tilde{\boldsymbol{\theta}}_{\text{SNP}_0}, \widehat{\boldsymbol{\theta}}^{(1)}_{\text{SNP}_L})$ matrix in (11) with all entries close to 0.

Given the theoretical result in (10), the GH test is given by

$$\mathrm{GH} = (\hat{\boldsymbol{\theta}}_{\mathrm{SNP}_L} - \tilde{\boldsymbol{\theta}}_{\mathrm{SNP}_0})' \hat{S}(\tilde{\boldsymbol{\theta}}_{\mathrm{SNP}_0}, \hat{\boldsymbol{\theta}}^{(1)}_{\mathrm{SNP}_L})^{-1} (\hat{\boldsymbol{\theta}}_{\mathrm{SNP}_L} - \tilde{\boldsymbol{\theta}}_{\mathrm{SNP}_0}). \tag{13}$$

Under normality of the latent variable, the GH test is asymptotically distributed as a χ^2_{2p}, with $2p$ degrees of freedom, that is, the number of parameters in $\boldsymbol{\theta}$.

We consider a simpler version of (13) that does not involve the inversion of the matrix $\hat{S}(\tilde{\boldsymbol{\theta}}_{\mathrm{SNP}_0}, \hat{\boldsymbol{\theta}}^{(1)}_{\mathrm{SNP}_L})$ giving

$$\mathrm{GH}_T = (\hat{\boldsymbol{\theta}}_{\mathrm{SNP}_L} - \tilde{\boldsymbol{\theta}}_{\mathrm{SNP}_0})' I^{-1}_{2p} (\hat{\boldsymbol{\theta}}_{\mathrm{SNP}_L} - \tilde{\boldsymbol{\theta}}_{\mathrm{SNP}_0}), \tag{14}$$

where I_{2p} can be omitted from the above formula.

Following Yuan and Bentler (2010)

$$\mathrm{GH}_T = \sum_{l=1}^{d} \lambda_l \delta_l^2, \delta_l \sim N(0,1), \tag{15}$$

where d is the rank of $S(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0{}^{(1)})$ and $\lambda_1, \ldots, \lambda_d$ are its non-zero eigenvalues.

It is possible to approximate the distribution of GH_T using the moment matching method (Welch, 1938; Yuan & Bentler, 2010) as follows:

$$\mathrm{GH}_T \sim a\chi^2_b. \tag{16}$$

The quantities a and b are defined as

$$a = \frac{\sum_{l=1}^{d} \lambda_l^2}{\sum_{l=1}^{d} \lambda_l} \tag{17}$$

and

$$b = \frac{(\sum_{l=1}^{d} \lambda_l)^2}{\sum_{l=1}^{d} \lambda_l^2}. \tag{18}$$

As $S(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0{}^{(1)})$ can be consistently estimated by $\hat{S}(\tilde{\boldsymbol{\theta}}_{\mathrm{SNP}_0}, \hat{\boldsymbol{\theta}}^{(1)}_{\mathrm{SNP}_L})$ defined in (11), a and b can be consistently estimated substituting $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ in (17) and (18), where d is rank of $\hat{S}$ and $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ are its non-zero eigenvalues. The approximation in (16) matches the first two moments of GH_T with those of $a\chi^2_b$ (16).

Similar to the statistics derived by Monroe (2021) and Li and Cai (2018), the GH_T test can be sensitive to the misspecification of either the latent variable distribution or the IRF. For this reason, correct IRFs for the items are assumed when testing the normality of the latent variable.

To study the sensitivity of the proposed GH_T test to misspecification of the IRF, we run a small-scale simulation and provide the results in Section S2.3 of Data S1.

4 | AVAILABLE TEST STATISTICS FOR TESTING LATENT VARIABLE DISTRIBUTION FIT AND THE OVERALL FIT

In addition to the proposed test statistic, we review three test statistics for testing latent variable distribution fit and one for the overall goodness-of-fit of the model considered each time.

An LR test statistic for nested models can be used since the SNP_L and SNP_0 models are nested (Wilks, 1938). The SNP_L model, when $\varphi_1 = \cdots = \varphi_L = \frac{\pi}{2}$, reduces to the SNP_0 model (Irincheeva et al., 2012). For the computation of the LR test, the SNP_0 model needs to be estimated using ML

instead of pairwise likelihood to obtain a comparable value of the log-likelihood function with that of the SNP_L model.

Let us denote by $\boldsymbol{\varphi}' = (\varphi_1, \cdots, \varphi_L)$ and by $\mathbf{c}' = (\frac{\pi}{2}, \cdots, \frac{\pi}{2})$. The null and alternative hypotheses can be formulated as follows:

$$H_0: \boldsymbol{\varphi} = \mathbf{c} \text{ vs } H_1: \boldsymbol{\varphi} \neq \mathbf{c}. \tag{19}$$

The test statistic is defined as

$$\text{LR} = 2\{l_{\text{SNP}_L}(\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(1)}) - l_{\text{SNP}_0}(\mathbf{y}, \widehat{\boldsymbol{\theta}})\}, \tag{20}$$

where $l_{\text{SNP}_L}(\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(1)})$ and $l_{\text{SNP}_0}(\mathbf{y}, \widehat{\boldsymbol{\theta}})$ are the log-likelihood functions of the SNP_L and SNP_0 models, respectively, evaluated at their maximum values. Under H_0, the LR test is asymptotically distributed as a χ^2_L, where L is the degree of the polynomial. For a fixed sample size, the number of empty cells in a frequency table increases with the number of binary items. In this case, the distribution of the LR test statistic is not well approximated by the chi-square distribution.

Li and Cai (2018) proposed a Pearson-type statistic based on summed scores given by:

$$\overline{X}^2 = n \sum_{s=0}^{S-1} \frac{(\overline{p}_s - \overline{\pi}_s)^2}{\overline{\pi}_s}, \tag{21}$$

where n is the sample size, $S = 1 + p$ represents the possible summed scores with p binary items, ranging from 0 to $S-1$. $\overline{p}_s$ and $\overline{\pi}_s$ denote the observed summed score proportions and the corresponding model-implied summed score probabilities computed under an IRT model, respectively, for $s = 0, \cdots, S-1$.

Li and Cai (2018) conjectured that under a wide variety of conditions, the tail-area probabilities of $\overline{X}^2$ can be well approximated by a χ^2 distribution with $S-1-2$ degrees of freedom under the null hypothesis that the latent variable distribution is correctly specified in the IRT model. The authors also proposed a moment-adjusted version of this test.

Monroe (2021) proposed a test statistic based on posterior residuals written as:

$$R(z) = \frac{\bar{f}(z|\mathbf{y}, \widehat{\boldsymbol{\theta}}) - d(z, \widehat{\boldsymbol{\theta}})}{s(z)}, \tag{22}$$

where $\widehat{\boldsymbol{\theta}}$ denotes the ML estimates of the free parameters of the IRT model, $\bar{f}(z|\mathbf{y}, \widehat{\boldsymbol{\theta}}) = n^{-1} \sum_{i=1}^{n} f(z|\mathbf{y}_i, \widehat{\boldsymbol{\theta}})$ denotes the sample average of the posterior distribution, $d(z, \widehat{\boldsymbol{\theta}})$ is the ML estimate of the distribution of the latent variable and $s(z)$ is the corresponding standard error estimate. If $d(z, \boldsymbol{\theta})$ is assumed to be standard normal, then $R(z)$ compares the posterior average, $\bar{f}(z|\mathbf{y}, \widehat{\boldsymbol{\theta}})$, with the standard normal distribution. A version of this test statistic for overall latent variable distribution fit, denoted by R_B, is defined as the maximum absolute of $R(z)$ and compared to a Bonferroni-corrected critical value.

We also review one more test statistic for the overall fit of the SNP_0 model proposed by Maydeu-Olivares and Joe (2005). Although not developed for testing latent distribution fit, it provides an overall goodness-of-fit test statistic that can work well under sparseness. 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, which 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 & Joe, 2005, 2006). Under the null hypothesis, we test that the SNP_0 model holds. The hypotheses H_0 and H_1 can be formalized as follows:

$$H_0: \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}) \text{ vs } H_1: \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta}), \tag{23}$$

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

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

$$M_2 = n\hat{\mathbf{e}}_2' \hat{\boldsymbol{U}}_2 \hat{\mathbf{e}}_2. \tag{24}$$

The vector $\hat{\mathbf{e}}$ includes the univariate and bivariate residuals while the matrix $\hat{\boldsymbol{U}}_2$ depends on a transformation matrix and the Jacobian matrix of the cell probabilities with respect to the items intercept and slope parameter (more details can be found in (Maydeu-Olivares & Joe, 2005, 2006). 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 LR, $\overline{X}^2$ and M_2 test statistics under normality and non-normality of the latent variable distribution. However, R_B is not available in commercial software and has therefore been reported only in the real data application in Section 7.1, where it was previously computed.

5 | MODEL SELECTION CRITERIA

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

The AIC is (Akaike, 1974):

$$\text{AIC} = -2l(\mathbf{y}, \hat{\boldsymbol{\theta}}^{(1)}) + 2k, \tag{25}$$

where $l(\mathbf{y}, \hat{\boldsymbol{\theta}}^{(1)})$ is the maximum value of the log-likelihood function of the SNP_L model and k is the number of free mode parameters.

The BIC is (Schwarz, 1978):

$$\text{BIC} = -2l(\mathbf{y}, \hat{\boldsymbol{\theta}}^{(1)}) + k\ln n, \tag{26}$$

where n is the sample size.

The HQ is (Hannan & Quinn, 1979):

$$\text{HQ} = -2l(\mathbf{y}, \hat{\boldsymbol{\theta}}^{(1)}) + 2k\ln\ln n. \tag{27}$$

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

When $L = 0$, $\hat{\boldsymbol{\theta}}^{(1)}$ equals $\hat{\boldsymbol{\theta}}$ in the aforementioned formulas.

6 | SIMULATION STUDY

6.1 | Design

In this section, we study the performance of the GH_T test to assess the non-normality of the latent variable distribution, and we compare its performance with the LR, $\overline{X}^2$ and M_2 test statistics. Moreover, the BIC, AIC and HQ criteria were computed for all simulation scenarios.

We have considered five scenarios (SC) corresponding to five distribution assumptions for the latent variable z in the data-generating models. The unidimensional latent trait model is

$$\text{logit}(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i \quad i = 1, \ldots, n \quad j = 1, 2, \ldots, p. \tag{28}$$

Item intercepts, α_{0j}, have been randomly chosen from the interval [−.8; 1.12] while the item slopes, α_{1j}, from the interval [.5; 1.5]. To study the Type I error rates of the above four test statistics, we have considered the following scenario:

A $z \sim N(0, 1)$.

For the power, we have considered the following cases of two bimodal distributions and two skewed distributions:

B $z \sim .7N(-1, .7) + .3N(1, .8)$,

where z has an overall mean of −.40 and a variance of 1.38.

C $z \sim .1N(-2, .25) + .9N(2, 1)$,

where z has an overall mean of 1.6 and a variance of 2.37.

D $Z \sim \text{SN}(\mu = -2.5, \sigma = 2, \lambda = 5)$,

where z has a mean of −.93 and a variance of 1.55.

E $Z \sim \text{SN}(\mu = -2.5, \sigma = 2, \lambda = 10)$,

where z has a mean of −.91 and a variance of 1.47.

The degree of deviation from the normal distribution is greater in scenario C than in scenario B, and similarly, the skew-normal distribution in scenario E exhibits a greater deviation from the normal distribution than the one in scenario D.

In the simulations, two versions of the GH_T test were considered. The first version, named GH_{T1}, is based on the SNP_0 and SNP_1 models. The SNP_1 model was chosen because it can approximate bimodal distributions well, present in scenarios B and C.

The SNP_1 model has been optimized in R using the 'nlminb' function, which maximizes directly through the analytically computed gradient and Hessian matrix. In the case of the SNP_1 model, the initial values for the parameters α_{0j} and α_{1j} used in the optimization process are the ML parameter estimates obtained with the SNP_0 model. For the φ_1 parameter, we have sampled 10 initial values from a sequence of values equally spaced by .1 within the domain of φ_1 which is the interval $[-\frac{\pi}{2}; \frac{\pi}{2}]$. From the estimated SNP_1 models in each data replication, we selected the one corresponding to the maximum value of the quasi-log-likelihood function. The GH_{T1} test has been evaluated under all scenarios to assess its performance.

In scenarios D and E, a second version of GH_T called GH_{T2} has been considered in addition to GH_{T1}. GH_{T2} is based on the SNP_0 and SNP_2 models, suitable for capturing highly skewed cases when $L = 2$, as shown in Figure 2. The SNP_2 model optimization uses the direct maximization function 'nlminb' in R. However, it is susceptible to the initial values chosen. To overcome this, the true parameter values

have been used as initial values for the α_{0j} and α_{1j} parameters, and the initial values for the φ_1 and φ_2 parameters have been set to 0.7 and 1, respectively, corresponding to a skew-normal case. To compute the GH_{T1} and GH_{T2} test statistics, the SNP_0 model has been estimated using the maximum PL method through the R function 'optim'. The Hessian, cross-product matrices and the matrix in formula (12) involved in the GH_{T1} and GH_{T2} tests have been computed numerically with the 'NumDeriv' R package.

To compute the values of the information criteria and the LR test, the SNP_0 model has been estimated using ML in R via the 'optim' function. For the M_2 test, we also used ML estimation. We estimated the 2PL model with the expectation-maximization (E-M) algorithm with the 'mirt' function in R. Then, we applied the 'M2' function to calculate the test statistic.

We used flexMIRT (Cai, 2017) to compute $\overline{X}^2$.

The Type I error rates and power of the GH_{T1}, GH_{T2}, LR, $\overline{X}^2$ and M_2 have been computed using the following formula: $\widehat{p} = \sum_{l=1}^{N_v} \frac{I(T_l \geq c)}{N_v}$. Here, N_v represents the number of valid replications out of the total, I denotes an indicator function and T_l is the test statistic value evaluated in the l-th replication. The critical value c corresponds to the theoretical asymptotic value, specifically the $(1-\alpha)$th percentile of the $\hat{a}\chi^2_{\hat{b}}$ distribution for the GH_T test. The values $\hat{a}$ and $\hat{b}$ are computed as in (17) and (18). For the M_2 test, the critical value is associated with the χ^2_m distribution, where m is equal to $\frac{p(p+1)}{2} - 2p$. For $\overline{X}^2$, the critical value is associated with the χ^2_{S-1-2} distribution.

To compare the performance of two models, SNP_1 and SNP_0, we use the LR test. As SNP_1 has an additional parameter compared to SNP_0, we use c as the theoretical asymptotic critical value corresponding to the $(1-\alpha)$th percentile of the χ^2_1 distribution. This specific test is called LR_1. To compare SNP_0 with SNP_2, we use the LR test with two degrees of freedom, denoted as LR_2. A confidence interval (CI) of each rate $\widehat{p}$ is computed as $\widehat{p} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\alpha(1-\alpha)}{N_v}}$.

Next, the percentages of times the AIC, BIC and HQ criteria select the SNP_1 over SNP_0 have been computed as $\widehat{P} = \sum_{l=1}^{N_v} \frac{I(IC_{SNP_1} < IC_{SNP_0})}{N_v} 100$, where IC indicates the AIC, BIC and HQ criteria. Similarly, we computed percentages for selecting the SNP_2 model over the SNP_0 model using information criteria.

We have conducted simulations for scenarios A, B and C with the following conditions: number of items ($p = 10, 20$), sample size ($n = 500, 1000, 5000$), test statistics (GH_{T1}, LR_1, $\overline{X}^2$, M_2) and information criteria (AIC, BIC, HQ). For scenarios D and E, the simulation conditions are the number of items ($p = 10, 20$), sample size ($n = 500, 1000$), test statistics (GH_{T1}, GH_{T2}, LR_1, LR_2, $\overline{X}^2$, M_2) and information criteria (AIC, BIC, HQ). We have considered two nominal levels of α, $\alpha = .05, .01$ and $R = 500$ replications for all scenarios.

Preliminary results from Guastadisegni et al. (2023) indicate that the GH_{T1} test performs well under scenarios A and C, particularly for large sample sizes, with reasonable rates of Type I error and power.

Additional results that include the bias for the parameter estimates of the model SNP_0, SNP_1 and SNP_2 have been reported in Section S2.1 of Data S1. When the mean of the true latent variable differs from 0, and the variance differs from 1, as in scenarios B, C, D and E, to compute the parameter bias, the true α_{00} and α_{01} parameters have been rescaled accordingly to formulas (S12) and (S13) in Section S1.3 of Data S1, replacing $\dot{\alpha}_{0j}$ with the true intercept, $\dot{\alpha}_{1j}$ with the true slope, $\tilde{E}(Z)$ 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 by Irincheeva (2011).

6.2 | Results

Table 1 reports the Type I error rates of the GH_{T1}, LR_1, $\overline{X}^2$ and M_2 for scenario A.

The GH_{T1} and $\overline{X}^2$ tests exhibit good Type I error rates under most conditions. However, GH_{T1} is more conservative than expected, but only when the sample size is small, and there are 10 or 20 items with $\alpha = .05$. The $\overline{X}^2$ test has lower than expected Type I error rates with a small sample size, 10 items and $\alpha = .05$. M_2

TABLE 1 Scenario A: Type I error rates of the GH_{T1}, LR_1, $\overline{X}^2$ and M_2, $p = 10, 20$, $n = 500, 1000, 5000$.

		$\alpha = .05$				$\alpha = .01$			
p	n	GH_{T1}	M_2	LR_1	$\overline{X}^2$	GH_{T1}	M_2	LR_1	$\overline{X}^2$
10	500	**.018**	.066	**0**	**.024**	.002	.012	0	.006
	1000	.044	.046	**.002**	.038	.006	.01	0	.014
	5000	.056	.04	.042	.052	.02	.002	0	.012
20	500	**.026**	.036	**.006**	.056	.008	0	0	.01
	1000	.042	.054	.054	.06	.008	.006	0	.012
	5000	.042	.05	**.264**	.052	.01	.016	**.094**	.018

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

TABLE 2 Scenario A: Percentages of times AIC, BIC and HQ select SNP_0 over SNP_1, $p = 10, 20$, $n = 500, 1000, 5000$.

p	n	AIC (%)	BIC (%)	HQ (%)
10	500	97	100	99.8
	1000	93.2	100	98.8
	5000	90.6	100	96.6
20	500	87.8	99.8	99.2
	1000	78.4	100	94.6
	5000	54.4	97	78.4

performs well regarding Type I error rates for both values of α, number of items and sample sizes. The LR_1 test performs the worst of the four tests. When there are 10 items and the sample size is 500 or 1000, or when there are 20 items, the sample size is 500, and $\alpha = .05$ rejects more than it should. Moreover, for every level of α considered, it has seriously inflated Type I error rates with 20 items and $n = 5000$.

Table 2 shows the percentages of times AIC, BIC and HQ select SNP_0 over SNP_1 for scenario A.

Among the three information criteria considered, the AIC has the worst performance. This is particularly noticeable with 20 items and all sample sizes, where it selects the SNP_0 model from 54% to 88% of times. On the contrary, the BIC performs the best, selecting the SNP_0 model almost every time under all conditions. The HQ performs moderately well in selecting the SNP_0 model, with its performance lying between that of the AIC and BIC criteria.

Table 3 presents the empirical power of the GH_{T1}, LR_1, $\overline{X}^2$ and M_2 for scenarios B and C.

Regarding the four tests and both levels of α considered, it was found that GH_{T1} consistently exhibits higher power than the other tests under scenario B for small sample sizes. Under scenario C, and generally as the sample size increases, GH_{T1} and $\overline{X}^2$ tend to have very similar power, which is consistently higher than LR_1. Overall, the power of the GH_{T1}, LR_1 and $\overline{X}^2$ tests increases as both the sample size and the number of items increase, and as the true latent variable distribution significantly diverges from the normal distribution. On the contrary, the M_2 test has very little or no power to detect non-normality of the latent variable distribution under both scenarios considered.

Table 4 shows the percentages of times AIC, BIC and HQ select SNP_1 over SNP_0 for scenarios B and C.

In all cases, AIC performs best when the sample size is 500 or 1000, with the highest percentage of selecting the SNP_1 model. On the contrary, BIC has the poorest performance with sample size of 500 or 1000, especially when it comes to selecting the SNP_1 model under scenario B and a sample size of 500, where it only selects it around 47% of the time. HQ performs better than BIC but not as well as AIC in selecting the SNP_1 model under all scenarios for small sample sizes.

TABLE 3 Scenarios B and C: Empirical power of the GH_{T1}, LR_1, $\overline{X}^2$ and M_2, $p = 10, 20$, $n = 500, 1000, 5000$.

			$\alpha = .05$				$\alpha = .01$			
SC	***p***	***n***	GH_{T1}	M_2	LR_1	$\overline{X}^2$	GH_{T1}	M_2	LR_1	$\overline{X}^2$
B	10	500	.612	.03	.572	.508	.536	.004	.462	.282
		1000	.868	.028	.792	.866	.796	.006	.558	.672
		5000	1	.052	.936	1	1	.014	.93	1
	20	500	.886	.022	.596	.728	.76	.006	.428	.472
		1000	.984	.036	.63	.984	.96	.002	.60	.928
		5000	.996	.034	.678	1	.996	.004	.666	1
C	10	500	.924	.01	.912	.992	.852	.002	.8	.976
		1000	.998	.006	.968	1	.994	0	.96	1
		5000	1	.028	.972	1	1	.004	.972	1
	20	500	.988	0	.782	1	.95	0	.72	1
		1000	.996	0	.822	1	.994	0	.756	1
		5000	.998	.004	.922	1	.998	0	.892	1

TABLE 4 Scenarios B and C: Percentages of times AIC, BIC and HQ select SNP_1 over SNP_0, $p = 10, 20$, $n = 500, 1000, 5000$.

SC	***p***	***n***	**AIC (%)**	**BIC (%)**	**HQ (%)**
B	10	500	75.6	47.6	58.8
		1000	85.2	55.4	78.8
		5000	94	93	93.4
	20	500	66.2	47	60.2
		1000	65.2	60	63
		5000	70	65.4	67.6
C	10	500	95.6	80.6	91.8
		1000	97	95.6	96.8
		5000	97.6	97.2	97.2
	20	500	85.8	73	79.4
		1000	87.6	75.4	82
		5000	95	87.2	92

Table 5 reports the empirical power of the GH_{T1}, LR_1, GH_{T2}, LR_2, M_2 and $\overline{X}^2$ for scenarios D and E.

Under scenarios D and E, GH_{T1} and GH_{T2} generally exhibit very similar power across all conditions. For small sample sizes, the power of these tests is consistently higher than that of $\overline{X}^2$. For large sample sizes, GH_{T1}, GH_{T2} and $\overline{X}^2$ demonstrate similar power, and it increases as the skew-normal distribution becomes more extreme. LR_1 has low power under all conditions, while LR_2 exhibits the highest power among all scenarios. Although increasing the degree of the polynomial can enhance the LR test performance, the SNP_2 method requires accurate initial values for the parameters to yield good results, making it challenging to use in practice. As observed in scenarios B and C, the M_2 test has very low or no power in detecting misspecifications of the latent variable distribution.

Table 6 shows the percentages of times AIC, BIC and HQ select SNP_1 over SNP_0 and SNP_2 over SNP_0 for scenarios D and E.

TABLE 5 Scenarios D and E: Empirical power of the GH_{T1}, LR_1, GH_{T2}, LR_2, M_2 and $\overline{X}^2$ tests, $p = 10, 20$, $n = 500, 1000$.

			$\alpha = .05$						$\alpha = .01$					
SC	p	n	GH_{T1}	LR_1	GH_{T2}	LR_2	M_2	$\overline{X}^2$	GH_{T1}	LR_1	GH_{T2}	LR_2	M_2	$\overline{X}^2$
D	10	500	.554	.374	.492	.7	.024	.39	.416	.27	.338	.462	.004	.186
		1000	.764	.588	.722	.956	.028	.742	.636	.264	.608	.85	.008	.504
	20	500	.786	.358	.89	.972	.026	.69	.576	.196	.762	.876	.008	.484
		1000	.962	.38	.992	1	.02	.972	.924	.348	.972	1	.002	.898
E	10	500	.588	.43	.562	.818	.028	.548	.458	.334	.404	.62	.01	.282
		1000	.894	.736	.814	.98	.036	.844	.834	.448	.688	.92	.014	.688
	20	500	.874	.45	.916	.99	.022	.808	.71	.282	.836	.946	.006	.626
		1000	.986	.474	1	1	.026	.996	.948	.458	.998	1	.026	.996

TABLE 6 Scenarios D and E: Percentages of times AIC, BIC and HQ select SNP_1 over SNP_0 and SNP_2 over SNP_0, $p = 10, 20$, $n = 500, 1000$.

			SNP_1 over SNP_0			SNP_2 over SNP_0		
SC	p	n	AIC (%)	BIC (%)	HQ (%)	AIC (%)	BIC (%)	HQ (%)
D	10	500	58.4	28.4	39	87	26.2	60.4
		1000	70.4	24.4	58.4	99	61.8	90
	20	500	44.8	22	36.6	98.8	75.4	94
		1000	41.4	34.6	38	100	98.2	100
E	10	500	64.6	34.4	44.2	91.6	42.2	72.6
		1000	81.2	44	73.2	99	76	95
	20	500	51.2	31	45.2	99.6	84.4	97.8
		1000	49.6	45.6	47.4	100	99.6	100

In both scenarios and with different sample sizes, none of the criteria were effective in choosing between SNP_1 and SNP_0. However, because the true latent variables are skewed and SNP_2 better approximates this type of distribution, all the information criteria performed better when selecting between SNP_0 and SNP_2. Among the criteria, AIC showed the best performance, while BIC performed the worst. Overall, based on the results in Tables 2, 4 and 6, it was not possible to determine which criterion performed the best under normality and non-normality of the latent variable.

To assess the performance of the GH_T test in the presence of misspecification of the IRF, an additional scenario, referred to as **F**, was examined. Details of the simulation conditions and results are in Section S2.3 of Data S1. The results indicate that GH_{T1} has lower power than $\overline{X}^2$. This suggests that if the GH_{T1} test rejects the null hypothesis, it is more likely due to misspecification in the latent variable distribution rather than in the IRF specification.

7 | REAL DATA APPLICATIONS

7.1 | Grade 12 science assessment test

The Grade 12 science assessment test (SAT12) dataset can be found in the TESTFACT manual (Wood et al., 2003) and is also included in the R package mirt (Chalmers, 2012). It comprises 32 binary-scored items measuring chemistry, biology and physics knowledge. Previous analysis by Monroe (2021) revealed

that the 2PL model provided a poor fit, as indicated by the significant values of R_B and $\overline{X}^2$. To further analyse this dataset, the SNP_0 and SNP_1 models were fitted to the same dataset used by Monroe (2021), which included 572 complete cases.

Section S3.1 of Data S1 reports estimates for the SNP_0 and SNP_1 models. These estimates show some differences, but they are not too large. To choose between the classic 2PL model and a more complex one, we first assessed the fit of the SNP_0 model. As the data are sparse, with all 570 observed response patterns having expected frequencies <5, we examined residuals calculated from marginal frequencies. We have used the rule that residuals >4 indicate a bad fit of the correspondent pair or triplets of items (Bartholomew et al., 2011). Although the results are not presented in the tables, the SNP_0 model exhibits a poor fit for specific pairs and triplets of items. We computed information criteria to assess whether the SNP_1 model would better fit.

Table 7 reports the AIC, BIC and HQ criteria for the SNP_0 and SNP_1 models. Based on the information criteria selected, the SNP_1 model is preferred, indicating the non-normality of the latent variable.

Table 8 reports the value of the M_2, LR_1, GH_{T1}, $\overline{X}^2$ and R_B, the degrees of freedom (dof) and the associated p-values.

The values of R_B and $\overline{X}^2$ were obtained from Monroe (2021). The R_B indicates that the latent variable is not normally distributed. The test M_2 rejects the null hypothesis that SNP_0 fits the data well. The LR_1 suggests a better fit for the SNP_1 model. Both GH_{T1} and $\overline{X}^2$ reject the null hypothesis that the latent variable is normally distributed. As some of the parameter estimates are more extreme than those considered in the simulation study in Section 6, the reliability of the test results was evaluated through a simulation study that mimics the real dataset, as explained in Section S3.1 of Data S1. The findings indicate that all tests perform well in terms of Type I error rates and power, except for the M_2 test, which shows very low or no power in identifying non-normality in the latent variable distributions.

This data analysis did not consider the GH_{T2} and LR_2 tests as we already rejected the null hypothesis when considering only $L = 1$.

7.2 | The NLSF dataset

This study is based on data collected from the National Longitudinal Survey of Freshmen (NLSF), which is a project funded by the Mellon Foundation and the Atlantic Philanthropies (available at http://oprdata.princeton.edu/archive/restricted), designed by Douglas S. Massey and Camille Z. Charles. The NLSF aims to collect data to explain the underachievement of minority groups in higher education.

TABLE 7 SAT12: Information criteria for the SNP_0 and SNP_1 models.

	AIC	BIC	HQ
SNP_0	18,239.49	18,517.84	18,348.08
SNP_1	18,199.7	18,482.39	18,309.98

TABLE 8 SAT 12 data: M_2, LR_1, GH_{T1}, $\overline{X}^2$ and R_B test statistics and associated degrees of freedom and p-values.

Test	Value	dof	*p*-value
M_2	675.53	464	<.001
LR_1	41.79	1	<.001
GH_{T1}	40.775	6.80	<.001
$\overline{X}^2$	172.8	30	<.001
R_B	7.59	–	<.001/31

The data were 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 first-year white, black, Asian and Latino students entering selective colleges and universities. For this study, we have analysed only a part of the questionnaire that refers to 1999. Specifically, we have selected 21 binary items, whose descriptions are reported in Table 9. The first 9 items measure violence in the neighbourhood, while the remaining 12 measure violence in the school.

The initial set of observations contained 3924 responses, with possible answers being 'no', 'yes', 'don't know' and 'refused'. Responses including 'don't know' or 'refused' were excluded from the analysis. In addition, 'no' responses were coded as 0 and 'yes' responses as 1. The final dataset used for analysis contained 3860 responses.

A subset of 400 observations, along with the items listed in Table 9, was analysed by Cagnone and Viroli (2012), who fitted a latent trait model to the data, with latent variables distributed as a finite mixture of multivariate Gaussians, to achieve both dimension reduction and clustering simultaneously. Cagnone and Viroli (2012) found a good fit for the model with 2 factors distributed as a mixture of Gaussians, with the first 9 items listed in Table 9 loading highly on one factor, and the remaining 12 on the other factor.

In this work, we analyse the two item batteries separately. The results are reported in the following subsections.

7.2.1 | American students exposure to neighbourhood violence

This section presents the analysis results of the nine items that measure neighbourhood violence.

TABLE 9 NLSF data: Item description.

Item	Question
1	In your neighbourhood, before you were 10
	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?
10	In your grade school, did you see students fighting?
11	Students smoking?
12	Students cutting class?
13	Students cutting school?
14	Students verbally abusing teachers?
15	Did you see physical violence directed at teachers by students?
16	Vandalism of school or personal property?
17	Theft of school or personal property?
18	Students consuming alcohol?
19	Students taking illegal drugs?
20	Students carrying knives as weapons?
21	Students with guns?

Despite the SNP_0 and SNP_1 methods being on the same scale, we observe different parameter estimates with very extreme values, as reported in Section S3.2.1 of Data S1.

We evaluated the fit of the SNP_0 model, considering the data's sparsity.

Specifically, 174 of the 231 observed response patterns have expected frequencies of <5. Residuals calculated from marginal frequencies were examined, revealing that the SNP_0 model does not fit well for some pairs and triplets of items.

Table 10 reports the values of the AIC, BIC and HQ criteria for the SNP_0 and SNP_1 models.

The information criteria give conflicting results. Indeed, AIC and HQ select the SNP_1 model, while BIC the SNP_0 model.

Table 11 reports the value of the M_2, LR_1, GH_{T1} and $\overline{X}^2$ test statistics, the degrees of freedom (dof) and the associated *p*-values.

The test results show contradictions. Specifically, the M_2 test indicates that the SNP_0 model does not fit the data well. For the LR_1 and GH_{T1} tests, the null hypothesis – that the latent variable is normally distributed – is rejected. In contrast, the $\overline{X}^2$ test does not reject the null hypothesis of normality for the latent variable. To further investigate these contradictory results, a simulation study was conducted with nine items. The study used the SNP_0 estimates (intercepts and slopes) as the true values for generating data. These estimates deviate significantly from the true parameter values used in the simulation presented in Section 7. The simulation results, presented in Section S3.2.1 of Data S1, indicate that all tests, except for LR_1, perform well regarding Type I error rates and when the latent variable is generated from a mixture of normals. Limitations arise under skew-normal scenarios. Specifically, the $\overline{X}^2$ and M_2 tests exhibit almost no power. While the GH_{T1} test shows slightly better performance than $\overline{X}^2$, its power remains low. Moreover, the SNP_1 method encountered convergence issues in nearly 20% of the replications. The results of the LR_1 test are unreliable due to significantly inflated false positive rates. Upon examining the results in Table 11 and those in Data S1, we conjecture that the situation observed in the real data is similar to scenarios D and E, where the GH_{T1} test demonstrates somewhat higher power than the $\overline{X}^2$ test. Nevertheless, we cannot trust the test results as the overall power is very low. Additionally, this dataset presents a specific challenge not encountered in the other data analysed or in the other simulation studies presented in the paper and Data S1: approximately 33% of the loadings for the SNP_0 model, which are used as data-generating values in the simulation study, are large, exceeding a value of 3. This factor could contribute to instability in the results and impact the estimation process of the SNP_1 model. While the application results remain unresolved, they could serve as a starting point for further investigation, as discussed in Section 8.4.

TABLE 10 NLSF data, nine items: Information criteria for the SNP_0 and SNP_1 models.

	AIC	BIC	HQ
SNP_0	21,309.32	21,422.12	21,349.36
SNP_1	21,303.67	21,422.73	21,345.93

TABLE 11 NLSF data, nine items: M_2, LR_1 GH_{T1}, $\overline{X}^2$ test statistics and associated degrees of freedom and *p*-values.

Test	Value	dof	*p*-value
M_2	182.33	27	0
LR_1	7.66	1	.006
GH_{T1}	222.92	2.65	0
$\overline{X}^2$	11.8	7	.106

7.2.2 | American students exposure to school violence

This section presents the analysis results on the 12 items that measure school violence. Also, in this battery of items, it has been observed that the parameter estimates reported in Section S3.2.2 of Data S1 are different despite the methods being on the same scale, and they exhibit extreme values. To choose the best model for the data, we first evaluated the fit of the SNP_0 model.

The data are sparse, with 227 out of the 302 observed response patterns having expected frequencies <5. Inspection of the bivariate and trivariate residuals showed that the SNP_0 model does not fit well.

Table 12 reports the values of the AIC, BIC and HQ criteria for the SNP_0 and SNP_1 models.

All three information criteria indicate that the SNP_1 model fits better than the SNP_0 model.

Table 13 reports the value of the M_2, LR_1, GH_{T1} and $\overline{X}^2$ test statistics, the degrees of freedom (dof) and their associated p-values.

Based on the M_2 test results, the SNP_0 model does not fit the data well. However, it is essential to note that the M_2 test does not identify the specific reason for this poor fit. On the contrary, the LR_1, GH_{T1} and $\overline{X}^2$ tests reject the null hypothesis that the latent variable is normally distributed. For this particular set of items, the test statistics and information criteria yield consistent results. This is evident from an additional simulation study we conducted, which mimics the parameter estimates from the SNP_0 model fitted to the 12 items. The results can be found in Section S3.2.2 of Data S1. In summary, the performance of the test statistics varies significantly based on the scenario being considered. None of the test statistics exhibit inflated false positive rates, and all tests demonstrate high power when the latent variable is generated from an extreme mixture of normals, including the M_2 test. The power is notably low only when the latent variable is generated from a skew-normal distribution, irrespective of the level of skewness.

8 | DISCUSSION

In the following subsections, we provide a summary of the work done in this paper, present the main findings with a focus on comparisons between the proposed tests and others, discuss the limitations of the GH_T test and outline directions for future work.

TABLE 12 NLSF data, 12 items: Information criteria for the SNP_0 and SNP_1 models.

	AIC	BIC	HQ
SNP_0	30,322.3	30,472.5	30,375.64
SNP_1	30,219.13	30,375.59	30,274.69

TABLE 13 NLSF data, 12 items: M_2, LR_1, GH_{T1} and $\overline{X}^2$ and associated degrees of freedom and p-values.

Test	Value	dof	p-value
M_2	979.39	54	<.001
LR_1	105.17	1	<.001
GH_{T1}	127.69	2.53	<.001
$\overline{X}^2$	68.4	10	<.001

8.1 | Summary of work

We have expanded the utilization of the GH test to detect the non-normality of the latent variable distribution in a unidimensional IRT model for binary data. The GH test has been derived through the comparison of the PL estimator of the 2PL model for binary data with the ML estimator of the SNP-IRT model, which allows for a more flexible shape of the latent variable distribution. We have used a simpler version of the GH test, referred to as the GH_T test, which does not require the inversion of the covariance matrix. Its approximated distribution has been derived using the moment matching method. Two versions of GH_T have been considered. The first version is the GH_{T1} test, based on the SNP_1. Through the simulation study in Section 6, as well as through three real data applications and the simulation studies based on these real applications reported in Data S1, we have compared the performance of the GH_{T1} test with the M_2, LR_1 and $\overline{X}^2$ test statistics and computed three information criteria, such as AIC, BIC and HQ. We also evaluated the sensitivity of the tests mentioned above to misspecifications of the item response function, with results reported in Data S1. As the SNP density when $L = 1$ can only capture bimodal and slightly skew-normal distributions, we have used the SNP density with $L = 2$ because this parameterization enables the representation of highly skew-normal distributions for specific combinations of parameter values. Therefore, we employed the second version of GH_T based on the SNP_0 model, namely GH_{T2}, and the LR_2 test only in the context of skew-normal scenarios in the main simulation study of Section 6 and in the study based on the SAT12 dataset to evaluate the performance of these tests with many items.

8.2 | Main findings

In the main simulation study discussed in Section 6, we used a standard range of values (without extreme values) to generate the intercepts and slopes of the items. The results showed that the GH_{T1} test performs well in terms of Type I error rates under most conditions. Furthermore, it demonstrates the highest power for detecting non-normality in the latent variable distribution across various scenarios, particularly with small sample sizes. Specifically, GH_{T1} outperforms $\overline{X}^2$ in terms of power for small sample sizes in the majority of cases, while both tests exhibit similar power when sample sizes are large. The GH_{T1} test consistently exhibits higher power than the LR_1 test, which can produce inflated or deflated Type I error rates under specific conditions. Additionally, Irincheeva (2011) noted some limitations in using the LR test within the framework of SNP models. Furthermore, the GH_{T1} test outperforms the M_2 test in terms of power. While the M_2 performs well regarding Type I error rates, it shows very low or no power to detect non-normality in the latent variable distribution. The only exception to this occurs in extreme mixture of normal distributions, as shown in simulation studies utilizing the NLSF dataset (with 9 and 12 items) reported in Data S1. In this scenario, the M_2 test shows high power to detect this misspecification. Similar findings regarding the low power of the M_2 test to identify non-normality in the latent variable distribution have also been reported by Li and Cai (2018), Paek et al. (2019) and Monroe (2021). As pointed out by Li and Cai (2018), this phenomenon can be attributed to the fact that M_2 relies on the first- and second-order margins of the underlying contingency table. To identify any misfit in the latent variable distribution, it may be necessary to incorporate information from higher-order margins.

The GH_{T1} test proves to be more effective than the information criteria in this context, as none of the criteria consistently outperforms the others. The simulations indicate that it is challenging to determine which information criterion performs best, whether the latent variable is normally or non-normally distributed. The AIC often selects more complex models in the highest percentage of cases, regardless of the distribution of the latent variable. This tendency to favour models with more parameters can result in overfitting. On the contrary, the BIC performs best when the latent variable follows a normal distribution. This is because BIC applies a stronger penalty for the number of parameters compared to the AIC, making it more suitable for selecting parsimonious models. However, BIC performs poorly

under non-normal distributions for small sample sizes, as it tends to select models that are too simple and fail to capture the complexity of the data. The performance of the HQ criterion lies between that of AIC and BIC.

From the additional simulation studies reported in Data S1, we found that the GH_{T1} test is less sensitive to misspecifications in the item response function compared to $\overline{X}^2$, particularly with a small number of items and small sample sizes. Therefore, if the GH_{T1} test results in a rejection, it is more likely that the misspecification is related to the latent variable distribution rather than the item response functions.

The simulation study based on the SAT12 dataset demonstrated that the performance of both the GH_{T1} and GH_{T2} tests remains stable, even when the number of items increases and certain item parameter values become more extreme. Overall, the results for all tests and information criteria are consistent with those observed in the main simulation study discussed in Section 6. Finally, the GH_{T1} and GH_{T2} tests show similar performance under skew-normal scenarios.

8.3 | Limitations of the proposed test

We identified two limitations in the GH_{T1} and GH_{T2} tests discussed in this work mainly in the case of examples with large slopes. For the GH_{T1} test, issues arose particularly during the simulation studies based on the NLSF dataset mentioned in Data S1. Specifically, when data were generated with extreme values for item slopes under skew-normal scenarios, the power of the GH_{T1} test was notably low. Additionally, in simulations involving nine items, the SNP_1 model, which the test relies on, encountered convergence problems. It is also important to note that neither $\overline{X}^2$ nor the information criteria performed well under the skew-normal scenarios in the simulations based on the NLSF dataset, especially with nine items. Furthermore, both LR_1 and M_2 exhibited limitations in these particular simulation studies. The second limitation is on the GH_{T2} test that requires estimation of the SNP_2 model, which poses a challenge as accurate initial values are necessary for the optimization process to attain a reliable approximation of the true latent variable and minimize bias in parameter estimates compared to SNP_0. Also the LR_2 test, which shows high power under the skew-normal scenarios in the simulation studies, is affected by the same problem. As a result, applying the SNP_2 model and the tests based on it in real data analysis becomes impractical due to these demanding requirements.

8.4 | Future research

Further studies could address the issue of initial values in SNP estimation when $L > 1$, enhancing its applicability in practical contexts. The performance of the GH test, when implemented with higher-order polynomials, could also be assessed through simulations and real data analysis. Increasing the degree of the polynomial allows for greater flexibility in modelling the shape of the latent variable distribution, which can improve both information criteria and test performance.

The limitations identified from the NLSF data application, particularly concerning the nine items, may serve as a foundation for further investigation. The GH test could be applied using other estimation methods and models different from SNP to improve performance, especially in situations where slope parameter values are more extreme.

Additionally, evaluating the performance of the GH test in the IRT context could involve examining other types of model violations, such as local dependence or violations of the item characteristic function. In these cases, it is important to consider other estimation methods that provide consistent estimates under model misspecification for the application of the test.

AUTHOR CONTRIBUTIONS

Lucia Guastadisegni: software; writing – original draft; methodology; data curation; visualization. **Silvia Cagnone:** methodology; writing – review and editing; formal analysis; supervision;

conceptualization; data curation. **Irini Moustaki:** methodology; conceptualization; writing – review and editing; supervision; formal analysis; writing – original draft; validation. **Vassilis Vasdekis:** conceptualization; methodology; writing – review and editing; supervision; formal analysis.

ACKNOWLEDGEMENTS

"This study was funded by the European Union-NextGenerationEU, in the framework of the "GRINS-Growing Resilient, INclusive and Sustainable project" (PNRR-M4C2-I1.3- PE00000018-CUP J33C22002910001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them."

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest with respect to the research, authorship and/orpublication of this article.

DATA AVAILABILITY STATEMENT

The data NLSF that support the findings of this study are openly available in the Data Archive at the Office of Population Research, Princeton University at http://oprdata.princeton.edu/archive/restricted. The data SAT12 that support the findings of this study are openly available in R.

ORCID

Silvia Cagnone https://orcid.org/0000-0002-7111-2945

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

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.

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

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

Cai, L. (2017). *flexMIRT® version 3.65: Flexible multilevel multidimensional item analysis and test scoring [computer software]*. Vector Psychometric Group.

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

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.

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.

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.

Guastadisegni, L., Moustaki, I., Vasdekis, V., & Cagnone, S. (2023). Detecting latent variable non-normality through the generalized hausman test. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology* (pp. 107–118). Springer Nature Switzerland.

Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of the American Statistical Association*, *108*(504), 1435–1444.

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*. PhD thesis, 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.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.

Li, Z., & Cai, L. (2018). Summed score likelihood–based indices for testing latent variable distribution fit in item response theory. *Educational and Psychological Measurement*, *78*(5), 857–886.
Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*(1), 221–239.
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.
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.
Monroe, S. (2021). Testing latent variable distribution fit in IRT using posterior residuals. *Journal of Educational and Behavioral Statistics*, *46*(3), 374–398.
Monroe, S. L. (2014). *Multidimensional item factor analysis with semi-nonparametric latent densities*. PhD thesis, 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.
Paek, I., Xu, J., & Lin, Z. (2019). Detection rates of the m2 test for nonzero lower asymptotes under normal and nonnormal ability distributions in the applications of irt. *Applied Psychological Measurement*, *43*(1), 84–88.
Ramsay, J. O. (2000). Differential equation models for statistical functions. *Canadian Journal of Statistics*, *28*(2), 225–240.
Ranger, J., & Much, S. (2020). Analyzing the fit of IRT models with the Hausman test. *Frontiers in Psychology*, *11*, 149.
Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*(1), 83–90.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.
van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.
Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, *92*(1), 1–28.
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.
Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62.
Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. (2003). *Testfact 4 for windows: Test scoring, item statistics, and full-information item factor analysis*. Scientific Software International.
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.
Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301.
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.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Guastadisegni, L., Cagnone, S., Moustaki, I., & Vasdekis, V. (2024). The generalized Hausman test for detecting non-normality in the latent variable distribution of the two-parameter IRT model. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–23. https://doi.org/10.1111/bmsp.12379