

# Score-Based Tests of Differential Item Functioning via Pairwise Maximum Likelihood Estimation

Ting Wang      Carolin Strobl      Achim Zeileis      Edgar C. Merkle  
University of Missouri      University of Zurich      Universität Innsbruck      University of Missouri

---

## Abstract

Measurement invariance is a fundamental assumption in item response theory models, where the relationship between a latent construct (ability) and observed item responses is of interest. Violation of this assumption would render the scale misinterpreted or cause systematic bias against certain groups of persons. While a number of methods have been proposed to detect measurement invariance violations, they typically require advance definition of problematic item parameters and respondent grouping information. However, these pieces of information are typically unknown in practice. As an alternative, this paper focuses on a family of recently-proposed tests based on stochastic processes of casewise derivatives of the likelihood function (i.e., scores). These score-based tests only require estimation of the null model (when measurement invariance is assumed to hold), and they have been previously applied in factor-analytic, continuous data contexts as well as in models of the Rasch family. In this paper, we aim to extend these tests to two-parameter item response models, with strong emphasis on pairwise maximum likelihood. The tests' theoretical background and implementation are detailed, and the tests' abilities to identify problematic item parameters are studied via simulation. An empirical example illustrating the tests' use in practice is also provided.

*Keywords:* pairwise maximum likelihood, score-based test, item response theory, differential item functioning.

---

## 1. Introduction

A major topic of study in educational and psychological testing is measurement invariance, with violation of this assumption being called *differential item functioning* (DIF) in the item response literature (see, for example, Millsap 2012, for a review). If a set of items violates measurement invariance, then individuals with the same ability (“amount” of the latent variable) may systematically receive different scale scores. This is problematic because researchers might conclude group ability differences when, in reality, the differences arise from unfair items.

We can formally define measurement invariance in a general fashion via (Mellenbergh 1989):

$$f(\mathbf{y}_i|v_i, \boldsymbol{\theta}_i) = f(\mathbf{y}_i|\boldsymbol{\theta}_i), \quad (1)$$

where  $\mathbf{y}_i$  is a vector of observed variables for individual  $i$ ,  $\boldsymbol{\theta}_i$  is the latent variable vector for individual  $i$ , which can be viewed as a random variable generated from a normal or multivariate normal distribution with parameter  $\boldsymbol{\theta}$ ,  $v_i \in V$ , where  $V$  is the auxiliary variable

such as age, gender, ethnicity, etc., against which we are testing measurement invariance, and  $f(\cdot)$  is an assumed parametric distribution. In applying the measurement invariance definition to a parametric item response theory (IRT) framework, Equation (1) states that the relationship between the latent construct (ability)  $\theta_i$  and response  $\mathbf{y}_i$  (binary or ordinal) holds regardless of the value of  $V$ .

Under this definition, many procedures have been proposed to assess measurement invariance/DIF in IRT models, including the Mantel-Haenszel statistic (Holland and Thayer 1988), Raju’s Area approach (Raju 1988), logistic regression methods (Swaminathan and Rogers 1990; Van den Noortgate and De Boeck 2005), Lord’s Wald test (Lord 1980) and likelihood ratio test (Thissen, Steinberg, and Wainer 1988). Overviews can be found in Millsap and Everson (1993); Osterlind and Everson (2009); Magis, Béland, Tuerlinckx, and De Boeck (2010); Glas (2015). These methods focus on generally detecting the presence or absence of DIF. When a measurement invariance violation is detected, however, researchers are typically interested in “locating” the measurement invariance. As Millsap (2005) stated, locating the invariance violation is one of the major outstanding problems in the field. This locating problem can be divided into two aspects. One is to locate which item parameter violates the measurement invariance assumption. The other is to locate the point/level of the auxiliary variable ( $V$ ) at which the violation occurs. Unfortunately, this second aspect is often ignored because previous procedures require us to pre-define the reference and focal groups (based on  $V$ ).

Beyond the approaches described in the previous paragraph, Glas and colleagues have done seminal work applying the *Lagrange multiplier* test (see also Satorra 1989) to item response models, focusing on situations where pre-defined group information is available (i.e.,  $V$  is treated as categorical variable). Their work has included the traditional two-parameter Glas (1998), three-parameter Glas and Falcón (2003), and nominal response Glas (1999, 2010) models, with applications including computerized adaptive testing Glas (2009), country-specific DIF Glas and Jehangir (2014) and models of response time Glas and Van der Linden (2010). The main estimation framework in this line of research is marginal maximum likelihood Glas (2009); Bock and Schilling (1997); Schilling and Bock (2005), which is generally most popular in IRT applications.

A more general family of *score-based* or *Lagrange multiplier* tests has been recently proposed to address “locating” issues in factor models for continuous response data Merkle and Zeileis (2013); Merkle, Fan, and Zeileis (2014); Wang, Merkle, and Zeileis (2014), where the auxiliary variable  $V$  can be continuous, ordinal, or categorical. Additionally, Strobl, Kopf, and Zeileis (2015) applied related tests to Rasch models estimated via conditional ML in order to identify the violating point along a categorical or continuous auxiliary variable. Moreover, Strobl *et al.* (2015) applied the tests recursively to multiple auxiliary variables via a “Rasch trees” approach, highlighting the fact that the groups tested for DIF need not be specified in advance and can even be formed by interactions of several auxiliary variables. Unfortunately, the conditional ML framework is only applicable to models of the Rasch family. Penalized maximum likelihood has also been recently proposed to detect DIF Tutz and Schauberger (2015), but the work has also been confined to the Rasch model.

In this paper, we extend the score-based tests to more general IRT models in a unified way, using both pairwise and marginal maximum likelihood estimation. We focus on identifying problematic item parameters without pre-specifying reference and focal groups. This approach allows us to 1) detect DIF in various IRT models without additional computational burden;

and 2) detect DIF against ordinal auxiliary variables like socioeconomic status and age group, whose ordinal nature is often ignored in IRT literature. We first describe the two-parameter IRT model and its relationship to factor analysis, along with the score-based tests' application to IRT via pairwise maximum likelihood estimation. Next, we report on the results of two simulation studies designed to examine the tests' ability to locate problematic item parameters while simultaneously handling the issue of person impact. Next, we apply the tests to real data, studying the measurement invariance of a mathematics achievement test with respect to socioeconomic status. Finally, we discuss test extensions and further IRT applications.

## 2. Model

In this study, we focus on binary data  $y_{ij}$ , where  $i$  represents individuals ( $i \in 1, \dots, n$ ) and  $j$  represents items ( $j \in 1, \dots, p$ ). There are two related approaches in the social science literature for analyzing these data: IRT and factor analysis. A two-parameter IRT model can be written as

$$y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad (2)$$

$$\text{logit}(p_{ij}) = \alpha_j \theta_i + \gamma_j, \quad (3)$$

$$\theta_i \sim N(\mu, \sigma^2), \quad (4)$$

where Equation (2) states that each person's response to each item ( $y_{ij}$ ) arises from a Bernoulli distribution with parameter  $p_{ij}$ . Then Equation (3) transforms  $p_{ij}$  to  $\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$ , which is a linear function of the person's ability  $\theta_i$  and the item parameters  $\gamma_j$  and  $\alpha_j$ . The alternative parameterization,  $\alpha_j(\theta_i - \gamma_j)$ , could also be used here. Finally, person ability  $\theta_i$  is described by hyperparameters  $\mu$  and  $\sigma^2$ , with these parameters commonly being fixed to 0 and 1, respectively, for identification. Instead of using the logit as the link function in Equation (3), we can alternatively use the inverse cumulative distribution function of the standard normal distribution  $\Phi^{-1}(\cdot)$  (the probit link function). In this case, Equation (3) could be written as  $p_{ij} = \Phi(\alpha_j \theta_i + \gamma_j)$ .

Use of the probit link function in the above model is equivalent to placing a factor analysis model on latent continuous variables  $\mathbf{y}^*$  Takane and De Leeuw (1987). In particular,

$$\mathbf{y}_i^* = \mathbf{\Lambda} \theta_i + \boldsymbol{\epsilon}, \quad (5)$$

where  $\mathbf{\Lambda}$  is a  $p \times 1$  factor loading vector, with components  $\lambda_1, \dots, \lambda_p$ ;  $\theta_i \sim N(0, 1)$ ; and  $\boldsymbol{\epsilon}$  is an error term, which follows the distribution  $N(\mathbf{0}, \mathbf{\Psi})$ . The matrix  $\mathbf{\Psi}$  is diagonal and defined as  $\mathbf{I} - \text{diag}(\mathbf{\Lambda} \mathbf{\Lambda}^\top)$ . The continuous response vector  $\mathbf{y}_i^*$  is composed by  $y_{ij}^*$  ( $j = 1, \dots, p$ ), with the observed binary data being obtained via

$$y_{ij} = \begin{cases} 1 & y_{ij}^* \geq \tau_j \\ 0 & y_{ij}^* < \tau_j. \end{cases} \quad (6)$$

Therefore, we can see that  $\lambda_j$  is similar to  $\alpha_j$  in Equation (3); they are both attached to the ability variable  $\theta_i$ . The error term  $\boldsymbol{\epsilon}$  is related to the probit link function that could be used in Equation (3). Finally, the threshold  $\tau_j$  corresponds to  $\gamma_j$ , which is related to item  $j$ 's difficulty.

No matter which link function is used, however, estimation of the two-parameter IRT model is not straightforward. The difficulty is caused by the person parameters  $\theta_i$ , which we generally avoid estimating (either by conditioning on them or integrating them out). Estimation methods that address this difficulty include conditional maximum likelihood (CML; e.g., Fischer and Molenaar 2012; De Ayala 2009), marginal maximum likelihood (MML; e.g., Thissen 1982) and pairwise maximum likelihood (PML; e.g., Katsikatsou, Moustaki, Yang-Wallentin, and Jöreskog 2012). We briefly describe PML below, which is the main focus of this paper (though we also consider MML).

### 3. Estimation

If we employ the factor analysis version of the model, the log-likelihood of individual  $i$ 's observed data  $\mathbf{y}_i$ , given the parameter vector  $\boldsymbol{\eta}$  (including  $\lambda$ ,  $\tau$ ), involves the integral

$$\ell(\boldsymbol{\eta}; \mathbf{y}_i) = \log \int_{\boldsymbol{\tau}} f(\mathbf{y}_i^* | \boldsymbol{\eta}) d\mathbf{y}_i^* \quad (7)$$

where  $\mathbf{y}_i^*$  is described as Equation (5), the distribution of  $\mathbf{y}_i^*$  with  $\theta_i$  marginalized out is denoted as  $f(\mathbf{y}_i^* | \boldsymbol{\eta})$  ( $p$  dimensional), which can be considered as a multivariate normal distribution following  $N(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})$ . The integration of the  $p$ -dimensional multivariate normal distribution over support  $\boldsymbol{\tau}$  is the difficult part, which does not have a closed form.

Katsikatsou *et al.* (2012) proposed that the likelihood function above can be approximated by:

$$p\ell(\boldsymbol{\eta}; \mathbf{y}_i) = \left\{ \sum_{j < k} \ell(\boldsymbol{\eta}; (y_{ij}, y_{ik})) \right\}, \quad (8)$$

$$= \left\{ \sum_{j < k} \left( \sum_{c_j=1}^2 \sum_{c_k=1}^2 \log \pi_{y_{ij}y_{ik}}^{(c_j c_k)}(\boldsymbol{\eta}) \right) \right\}, \quad (9)$$

where  $\sum_{j < k} \ell(\boldsymbol{\eta}; (y_{ij}, y_{ik}))$  is the log-likelihood associated with all pairs of items, which is a series of 2-way contingency tables;  $\pi_{y_{ij}y_{ik}}^{(c_j c_k)}(\boldsymbol{\eta})$  is the probability that individual  $i$  responds to item  $j$  and  $k$  with category  $c_j$  ( $c_j = 1, 2$ ) and  $c_k$  ( $c_k = 1, 2$ ) under the model, which is expressed as a function of pairwise integrals. (See Katsikatsou *et al.* 2012) for the explicit expression of  $\pi_{y_{ij}y_{ik}}^{(c_j c_k)}(\boldsymbol{\eta})$ , and note that the categories 1, 2 above represent responses of "0", "1" respectively in Equation (6).

Comparing Equation (7) with Equation (8), we can see that the  $p$ -dimensional integral is reduced to all possible pairwise ( $j < k$ ) integrals, which are bivariate normal distributions with closed form solution. This significantly reduces the computational complexity, which is a major advantage of PML.

### 4. Maximizing likelihood function

The model's log-likelihood function can be written as the sum of individual log-likelihoods

$$\ell(\boldsymbol{\eta}; \mathbf{y}_1, \dots, \mathbf{y}_n) = \sum_{i=1}^n \log f(\mathbf{y}_i | \boldsymbol{\eta}), \quad (10)$$

where the length of the parameter vector  $\boldsymbol{\eta}$  is  $q$ .

Maximizing the model's log-likelihood function is equivalent to solving the first order conditions

$$\sum_{i=1}^n \mathbf{s}(\hat{\boldsymbol{\eta}}; \mathbf{y}_i) = \mathbf{0}, \quad (11)$$

where

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \ell(\boldsymbol{\eta}; \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n). \quad (12)$$

and

$$\mathbf{s}(\hat{\boldsymbol{\eta}}; \mathbf{y}_i) = \left. \frac{\partial \ell(\mathbf{y}_i, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \right|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}} \quad (13)$$

$$= \left( \frac{\partial \ell(\boldsymbol{\eta}; \mathbf{y}_i)}{\partial \eta_1}, \dots, \frac{\partial \ell(\boldsymbol{\eta}; \mathbf{y}_i)}{\partial \eta_q} \right). \quad (14)$$

For the two-parameter IRT model analyzed in this paper, the log-likelihood function and consequently also the individual score function differs depending on the log-likelihood. We provide some detail on the PML score function below, with the MML score function being detailed in, e.g., [Glas \(1998\)](#).

Maximizing the log-likelihood function in Equation (8) over the parameter  $\boldsymbol{\eta}$ , we obtain the composite pairwise maximum likelihood estimator  $\hat{\boldsymbol{\eta}}_{\text{PML}}$ . Again, this is equivalent to solving for  $\boldsymbol{\eta}$  so that the sum of scores equals zero. The score vector of the pairwise likelihood for each individual can be decomposed in two blocks: the first derivative with respect to the factor loading  $\boldsymbol{\Lambda}$  and the first derivative with respect to the thresholds  $\boldsymbol{\tau}$ :

$$\mathbf{s}(\boldsymbol{\eta}; \mathbf{y}_i) = \left( \frac{\partial \left\{ \sum_{j < k} \ell(\boldsymbol{\eta}; (y_{ij}, y_{ik})) \right\}}{\partial \boldsymbol{\Lambda}}, \frac{\partial \left\{ \sum_{j < k} \ell(\boldsymbol{\eta}; (y_{ij}, y_{ik})) \right\}}{\partial \boldsymbol{\tau}} \right). \quad (15)$$

The elements of the score matrix are *analytic* solutions, requiring no approximation via quadrature. The derivatives are explicitly demonstrated in the Appendix of [Katsikatsou et al. \(2012\)](#).

It is easy to show that PML estimates are more easily obtained and less computationally intensive compared to traditional maximum likelihood estimation, e.g., marginal maximum likelihood (MML) estimation, which often involves quadrature or adaptive quadrature to approximate integrals [Schilling and Bock \(2005\)](#); [Katsikatsou et al. \(2012\)](#). Further, [Katsikatsou and Moustaki \(2016\)](#) have recently derived likelihood ratio tests for the PML framework, which in turn lead to expressions for pairwise AIC and BIC. Thus, we focus on PML in the simulations and analyses below, with similar results holding for MML as demonstrated in supplementary material. In the next section, we describe the scores' use in tests of measurement invariance.

## 5. Score-based tests of measurement invariance

Measurement invariance is usually studied in a hypothesis testing framework. We can write

the hypothesis very generally by assuming a potential observation-specific parameter vector  $\boldsymbol{\eta}_i$ . The null hypothesis of measurement invariance can then be expressed as all observations arising from a common set of population parameters  $\boldsymbol{\eta}_0$

$$H_0 : \boldsymbol{\eta}_i = \boldsymbol{\eta}_0 \quad (i = 1, \dots, n), \quad (16)$$

versus

$$H_1 : \boldsymbol{\eta}_i = \boldsymbol{\eta}(v_i) \quad (i = 1, \dots, n), \quad (17)$$

where  $\boldsymbol{\eta}(v_i)$  is typically an unknown function w.r.t.  $v_i$ . If the function is known, the alternative hypothesis can be expressed more specifically. For example, one function of particular interest involves  $V$  dividing individuals into two subgroups with different parameter vectors based on the cut point  $v$ :

$$H_1 : \boldsymbol{\eta}_i = \begin{cases} \boldsymbol{\eta}^{(A)} & v_i \leq v \\ \boldsymbol{\eta}^{(B)} & v_i > v. \end{cases} \quad (18)$$

For this hypothesis testing problem with known cut point  $v$ , the likelihood ratio test (LRT; [Thissen \*et al.\* 1988](#)) is most popular. The LRT compares two models, a full model and a reduced model. The full model is a multiple-group model with parameters free to vary across group A and group B, while the reduced model constrains some parameters to be equal across groups. The LRT statistic for cut point  $v$  can be expressed as

$$LR(v) = -2[\ell(\hat{\boldsymbol{\eta}}; \mathbf{y}_1, \dots, \mathbf{y}_n) - \{\ell(\hat{\boldsymbol{\eta}}^{(A)}; \mathbf{y}_1, \dots, \mathbf{y}_m) + \ell(\hat{\boldsymbol{\eta}}^{(B)}; \mathbf{y}_{m+1}, \dots, \mathbf{y}_n)\}], \quad (19)$$

where  $\ell$  represents the log-likelihood function,  $\hat{\boldsymbol{\eta}}^{(A)}$  is the MLE of  $\boldsymbol{\eta}^{(A)}$  based on  $\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ , for which  $v_i \leq v$  and  $\hat{\boldsymbol{\eta}}^{(B)}$  is the MLE of  $\boldsymbol{\eta}^{(B)}$  based on  $\{\mathbf{y}_{m+1}, \dots, \mathbf{y}_n\}$  for which  $v_i > v$ . This LRT statistic has an asymptotic  $\chi^2$  distribution with degrees of freedom equal to the number of parameters in  $\boldsymbol{\eta}$ .

However, when the grouping information is unknown, we can also compute  $LR(v)$  for each possible value of  $V$  in some interval  $[\underline{v}, \bar{v}]$ , obtaining a test statistic via:

$$\max_{v \in [\underline{v}, \bar{v}]} LR(v). \quad (20)$$

The asymptotic distribution of this maximum LR statistic is not  $\chi^2$ ; [Andrews \(1993\)](#) showed that, under the null hypothesis in (16), the statistic converges in distribution to some stochastic process. This result is also utilized in the score-based tests discussed below.

## 5.1. Test background

The score-based tests described here utilize the scores defined above, and they are based on theory showing that functions of the scores follow a stochastic process along an auxiliary variable  $V$ . Related descriptions can be found in [Zeileis and Hornik \(2007\)](#), [Merkle \*et al.\* \(2014\)](#), and [Wang \*et al.\* \(2014\)](#).

We can build the following intuition for the tests. We examine individuals' scores as we move from the smallest value of  $V$  to the largest. If there are no measurement invariance violations, the scores should fluctuate around zero. Conversely, the scores will systematically shift from zero when measurement invariance is violated.

To obtain formal test statistics, we define the cumulative score as

$$\mathbf{B}(t; \hat{\boldsymbol{\eta}}) = \hat{\mathbf{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{s}(\hat{\boldsymbol{\eta}}; \mathbf{y}_{(i)}) \quad (0 \leq t \leq 1), \quad (21)$$

where  $\mathbf{y}_{(i)}$  represents the observed data vector for  $i$ th-largest observation, with ordering determined by the auxiliary variable  $V$ .  $\hat{\mathbf{I}}$  denotes some estimate of the covariance matrix of the scores, which serves to decorrelate the fluctuation processes associated with individual model parameters;  $\lfloor nt \rfloor$  is the integer part of  $nt$  (i.e., a floor operator); and  $0 \leq t \leq 1$ . In a sample of size  $n$ ,  $\mathbf{B}(t; \hat{\boldsymbol{\eta}})$  changes at  $0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}$ . For  $t = 1$  the cumulative score vector always equals  $\mathbf{0}$ , as defined in Equation (11). We are specifically interested in how the cumulative score fluctuates as we move from  $t = 0$  to  $t = 1$ .

Along with the score vectors, we need an estimate of the score covariance matrix, which is shown in Equation (21) as  $\hat{\mathbf{I}}$ . For regular maximum likelihood estimation, the covariance matrix is equal to the information matrix. However, this identity does not hold for PML Katsikatsou *et al.* (2012). Therefore, instead of the information matrix, we use an estimate based on the outer product of scores  $\hat{\mathbf{I}} = (1/n) \sum_{i=1}^n \mathbf{s}(\hat{\boldsymbol{\eta}}; \mathbf{y}_{(i)}) \mathbf{s}(\hat{\boldsymbol{\eta}}; \mathbf{y}_{(i)})^\top$ .

Hjort and Koning (2002) showed that, under the null hypothesis from (16),  $\mathbf{B}(t; \hat{\boldsymbol{\eta}})$  converges in distribution to an independent Brownian bridge:

$$\mathbf{B}(\cdot; \hat{\boldsymbol{\eta}}) \xrightarrow{d} \mathbf{B}^0(\cdot), \quad (22)$$

where  $\mathbf{B}^0(\cdot)$ , is a  $q$ -dimensional Brownian bridge, and each column represents a unidimensional Brownian bridge associated with a single parameter.

Empirically, the  $\mathbf{B}(t; \boldsymbol{\eta})$  process can be described by an  $n \times q$  matrix, with each column following an independent Brownian bridge. The matrix row represents the ordered observations' cumulative score vector and the last row is zero as described by Equation (11). To obtain scalar test statistics, we summarize the empirical behavior of Equation (21) and compare it to the analogous scalar summary of the Brownian bridge. In the next section, we introduce various summaries of Equation (21) that can serve as test statistics.

## 5.2. Test statistics

After summarizing or aggregating the empirical cumulative score process via a scalar, the asymptotic distribution of the scalar can be obtained by applying the same summary to the asymptotic Brownian bridge. This allows us to obtain critical values and  $p$ -values. Various statistics have been proposed, with selection of a statistic being based on the plausible patterns of potential measurement invariance violations.

The simplest aggregation strategy is to reject measurement invariance if the largest component of the empirical cumulative score matrix is greater than a critical value. Based on the location of the detected component, we can easily identify the violating parameter and the value of  $V$  at which the violation occurs. Because this statistic is searching for the maximum over the parameters (columns of the empirical cumulative score matrix) and individuals (rows of the empirical cumulative score matrix), this statistic is called the ‘‘double maximum’’ ( $DM$ ).

$$DM = \max_{i=1, \dots, n} \max_{j=1, \dots, q} |\mathbf{B}(\hat{\boldsymbol{\eta}})_{ij}|. \quad (23)$$

However, the  $DM$  statistic is sub-optimal if many of the parameters change at the same value of  $V$  and/or there exist many (rather than only one) changing points in  $V$ , because this “wastes” power by only considering the maximum. In such cases, sums across parameters and individuals are more suitable. The Cramèr-von Mises ( $CvM$ ) statistic falls in this category,

$$CvM = n^{-1} \sum_{i=1, \dots, n} \sum_{j=1, \dots, q} \mathbf{B}(\hat{\boldsymbol{\eta}})_{ij}^2. \quad (24)$$

If we expect there is only one change point, but that change point affects multiple parameters, we can aggregate by summing over parameters, then taking the maximum over the individual interval (scaled by variance). This statistic is equivalent to obtaining the maximum of Lagrange multiplier statistics, and it can be formally written as

$$\max LM = \max_{i=\underline{i}, \bar{i}} \left\{ \frac{i}{n} \left( 1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1, \dots, q} \mathbf{B}(\hat{\boldsymbol{\eta}})_{ij}^2. \quad (25)$$

Note that this statistic is asymptotically equivalent to the  $\max LR$  mentioned before, in the same way that the traditional likelihood ratio test is asymptotically equivalent to the traditional Lagrange multiplier test.

Across the above statistics, the auxiliary variable  $V$  is assumed to be continuous. [Merkle et al. \(2014\)](#) introduced two modified statistics that could deal with ordinal  $V$ , which could include school grades or income levels. For an ordinal auxiliary variable with  $m$  levels, the modifications are based on  $t_l$  ( $l = 1, \dots, m - 1$ ), which are the empirical, cumulative proportions of individuals observed at the first  $m - 1$  levels. The modified statistics are then given by

$$WDM_o = \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left( 1 - \frac{i}{n} \right) \right\}^{-1/2} \max_{j=1, \dots, q} |\mathbf{B}(\hat{\boldsymbol{\eta}})_{ij}|, \quad (26)$$

$$\max LM_o = \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left( 1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1, \dots, q} \mathbf{B}(\hat{\boldsymbol{\eta}})_{ij}^2, \quad (27)$$

where  $i_l = \lfloor n \cdot t_l \rfloor$  ( $l = 1, \dots, m - 1$ ).

If the auxiliary variable  $V$  is only nominal/categorical, the empirical cumulative sums of scores can be used to obtain a Lagrange multiplier statistic by first summing scores within each of the  $m$  levels of the auxiliary variable, then summing the sums [Hjort and Koning \(2002\)](#). This test statistic can be formally written as

$$LM_{uo} = \sum_{l=1, \dots, m} \sum_{j=1, \dots, q} (\mathbf{B}(\hat{\boldsymbol{\eta}})_{lj} - \mathbf{B}(\hat{\boldsymbol{\eta}})_{l-1j})^2, \quad (28)$$

where  $\mathbf{B}(\hat{\boldsymbol{\eta}})_{i_0j} = 0$  for all  $j$ . This statistic is asymptotically equivalent to the usual likelihood ratio statistic, and it is advantageous over the LRT from (19) because it requires the estimation of only one model (the null model). We expect similar asymptotic equivalence results to hold in the PML framework, though this equivalence has not been investigated.

In the following sections, we apply these theoretical results to IRT models. We focus on the two-parameter model estimated via PML, where the  $\theta_i$  are assumed to arise from a normal distribution (and, as mentioned previously, MML results are included in the appendices).



## 6. Simulation 1

In this study, we aim to examine the tests' abilities to locate item parameters that violate measurement invariance. Consider a hypothetical battery of five items administered to students in several ordered groups (e.g.  $m = 8$ ), with the item responses being described by a traditional two-parameter model. Measurement invariance violations may occur in the item intercept or the item slope parameters (related to difficulty and discrimination, respectively). It is plausible that violations in an item's slope parameter influences the item's intercept parameter, or that one violating item influences the other items. Thus, the goal of Simulation 1 is to examine the extent to which the score-based tests attribute the measurement invariance violation to the correct item parameters.

### 6.1. Method

Data were generated from a two-parameter model (with probit link function) for a test with 5 items. A violation occurred in one of two places: the item 3 slope parameter ( $\alpha_3$ ) or intercept parameter ( $\gamma_3$ ). The fitted models matched the data-generating model, and parameter estimates were obtained by PML. Measurement invariance violations were tested in eight subsets of parameters: each item's intercept parameter (or slope parameter, depending on the location of the true violation), item 3's non-violating parameter ( $\gamma_3$  or  $\alpha_3$ ), all items' intercept parameters, and all items' slope parameters.

Power and type I error were examined across three sample sizes ( $n = 120, 480, 960$ ), three numbers of ordered groups ( $m = 4, 8, 12$ ) and 17 magnitudes of invariance violations. The measurement invariance violations occurred at level  $m/2+1$  of  $V$ : Students with  $V < (m/2+1)$  deviated from students with  $V \geq (m/2+1)$  by  $d$  times the parameters' asymptotic standard errors (scaled by  $\sqrt{n}$ ), with  $d = 0, 0.25, 0.5, \dots, 4$ .

For each combination of sample size ( $n$ )  $\times$  violation magnitude ( $d$ )  $\times$  violating parameter  $\times$  groups ( $m$ ), 5000 data sets were generated and tested. In all conditions, we maintained equal sample sizes in each subgroup of the categories  $m$ . Statistics from Equations (26) and (27) (both ordinal statistics) were examined, as was the statistic from (28) (categorical statistic, ignoring the ordering information). As mentioned previously, the latter statistic is asymptotically equivalent to the usual likelihood ratio test. Thus, this statistic provides information about the relative performance of the ordinal statistics vs. the LRT.

### 6.2. Results

Full simulation results for PML are presented in Figures 1 to 4 (similar results for MML are shown in supplementary material). Figures 1 and 2 compare different test statistics at a fixed value of  $n$ , while Figures 3 and 4 display a single test statistic across all values of  $n$ . Because items 1, 2, 4, and 5 display similar power curves in all conditions, we only show item 2's results.

Figure 1 demonstrates power curves (of sample size 960) as a function of violation magnitude in item 3's slope parameter  $\alpha_3$ , with the tested parameters changing across rows, the number of levels  $m$  of the ordinal variable  $V$  changing across columns, and lines reflecting different test statistics. In each panel, the x-axis represents the violation magnitude and the y-axis represents power. Figure 2 demonstrates similar power curves when the violating parameter is item 3's intercept parameter  $\gamma_3$ .

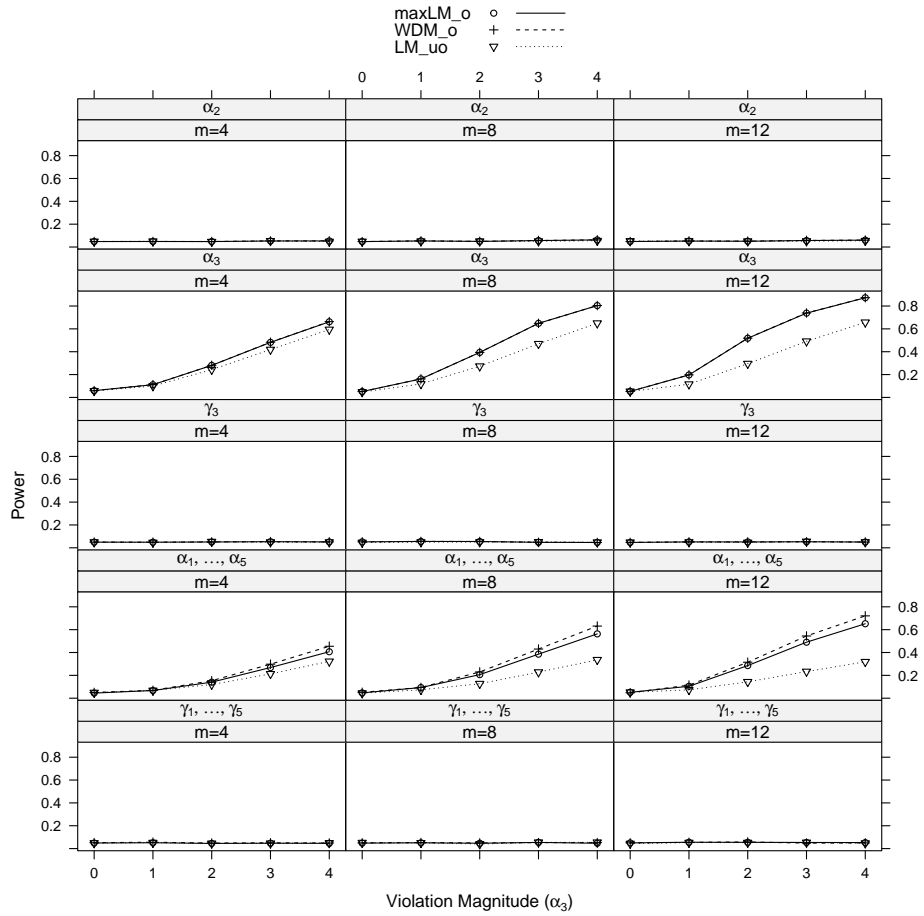


Figure 1: Simulation 1. Simulated power curves for  $\max LM_o$ ,  $WDM_o$ , and  $LM_{uo}$  across three levels of the ordinal variable  $m$  and measurement invariance violations of 0–4 standard errors (scaled by  $\sqrt{n}$ ), estimated by the PML two-parameter model. The parameter violating measurement invariance is  $\alpha_3$ .  $n = 960$ . Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable  $m$ .

These two graphs show that the ordinal statistics exhibit similar results, with the  $\max LM_{uo}$  statistic demonstrating lower power across all situations. This demonstrates the sensitivity of the ordinal statistics to invariance violations that are monotonic with  $V$ . In situations where only one parameter is tested,  $WDM_o$  and  $\max LM_o$  exhibit equivalent power curves. This is because these two statistics are equivalent when only one parameter is tested (see Merkle *et al.* 2014).

Figures 3 and 4 display similar power curves (of statistic  $WDM_o$ ), but the lines now reflect different sample sizes. Figure 3 demonstrates results when the violating parameter is  $\alpha_3$ , and Figure 4 displays the results when the violating parameter is  $\gamma_3$ .

From these figures, one generally observes that the tests isolate the parameter violating measurement invariance. Comparing Figure (1) to Figure (2), we can see the tests have somewhat higher power to detect measurement invariance violations in the intercept parameter as opposed to the slope parameter. This is because it is easier to detect violations in “main effects”

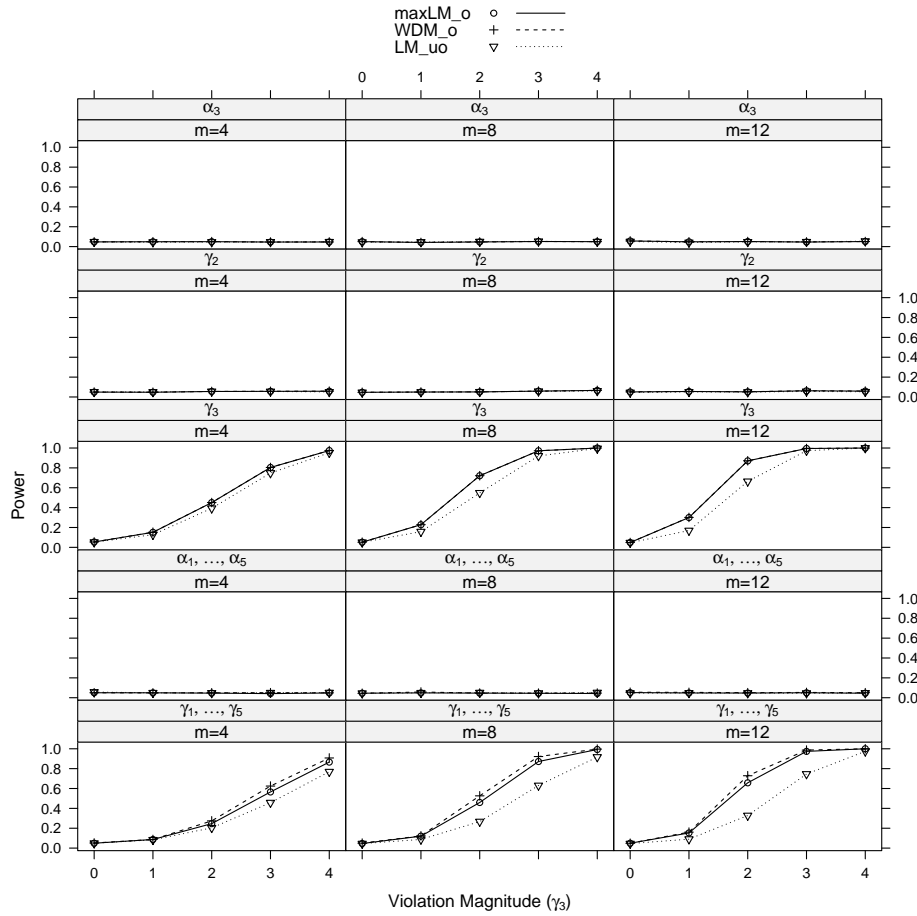


Figure 2: Simulation 1. Simulated power curves for  $\max LM_o$ ,  $WDM_o$ , and  $LM_{uo}$  across three levels of the ordinal variable  $m$  and measurement invariance violations of 0–4 standard errors (scaled by  $\sqrt{n}$ ), estimated by PML two-parameter model. The parameter violating measurement invariance is  $\gamma_3$ .  $n = 960$ . Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable  $m$ .

(we can see it as intercept  $\times 1$ ) than in “interactions” (slope  $\times$  person parameter  $\theta_i$ ). Any changes in an intercept parameter will influence every person equally whereas any changes in a slope parameter’s influence is moderated by each person’s ability  $\theta_i$ . Meanwhile, comparing Figure 3 and Figure 4, we can see that sample size has a much larger influence on power to detect violations in the slope parameter, as compared to the intercept parameter. This is related to the fact that the violation magnitudes were scaled by the square root of  $n$ , and the slope parameter is attached to the person parameter  $\theta_i$  which follows a distribution instead of a constant.

Finally, simultaneous tests of all slope parameters or of all intercept parameters resulted in decreased power, as compared to the situation where only the violating parameter is tested. This “dampening” phenomenon is more apparent for  $\max LM_o$  statistic, because it involves a sum across all tested parameters (see Equation (27)) whereas  $WDM_o$  only takes the maximum over parameters (see Equation (26)). However, the relative power advantage of using  $\max LM_o$

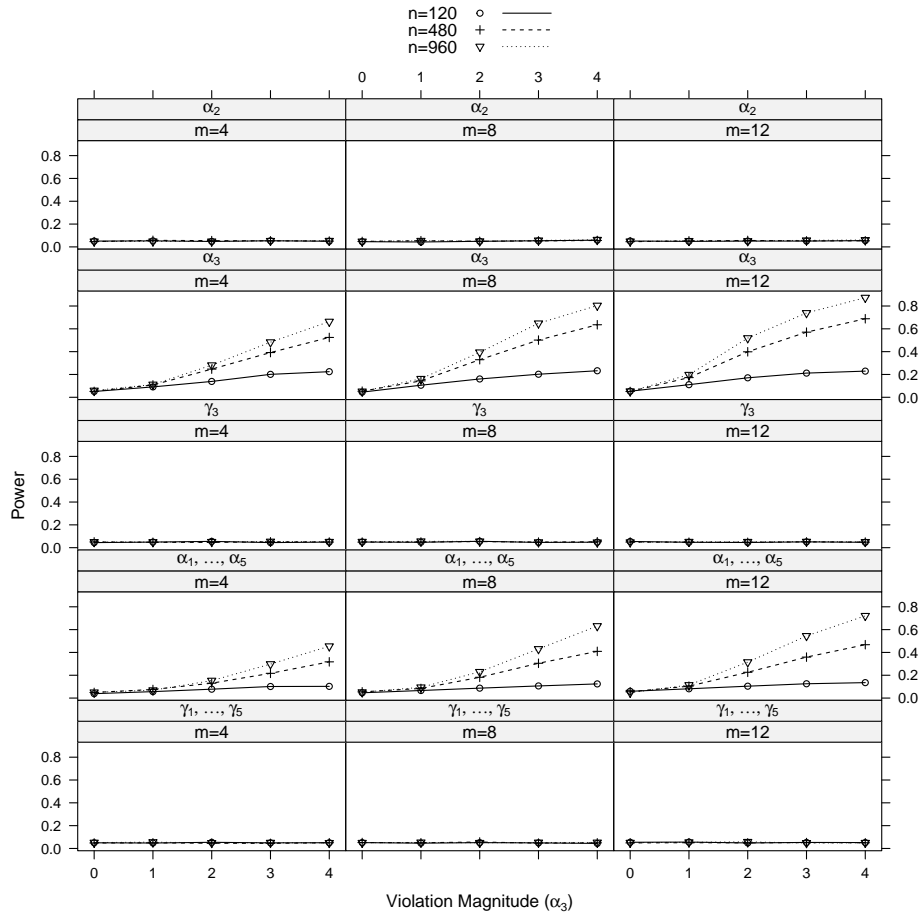


Figure 3: Simulation 1. Simulated power curves for sample sizes  $n = 120, 480$  and  $960$  of test statistic  $WDM_o$ , across three levels of the ordinal variable  $m$  and measurement invariance violations of 0–4 standard errors (scaled by  $\sqrt{n}$ ), estimated by PML two-parameter model. The parameter violating measurement invariance is  $\alpha_3$ . Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable  $m$ .

and  $WDM_o$  when testing multiple parameters depends on the number of parameters that actually violate invariance Merkle *et al.* (2014). In practice, we often test multiple parameters in the exploratory stage and, when we have no information about which parameter(s) might be problematic,  $\max LM_o$  has more power than  $WDM_o$  Merkle *et al.* (2014); Wang *et al.* (2014).

In summary, we found that the proposed tests can attribute measurement invariance violations to the correct parameter of a two-parameter item response model. While this can give practitioners some confidence in the tests, we did not examine the situation where person abilities differ across groups, which is often called “impact” in item response literature Fischer (1995b). We consider this situation in Simulation 2.

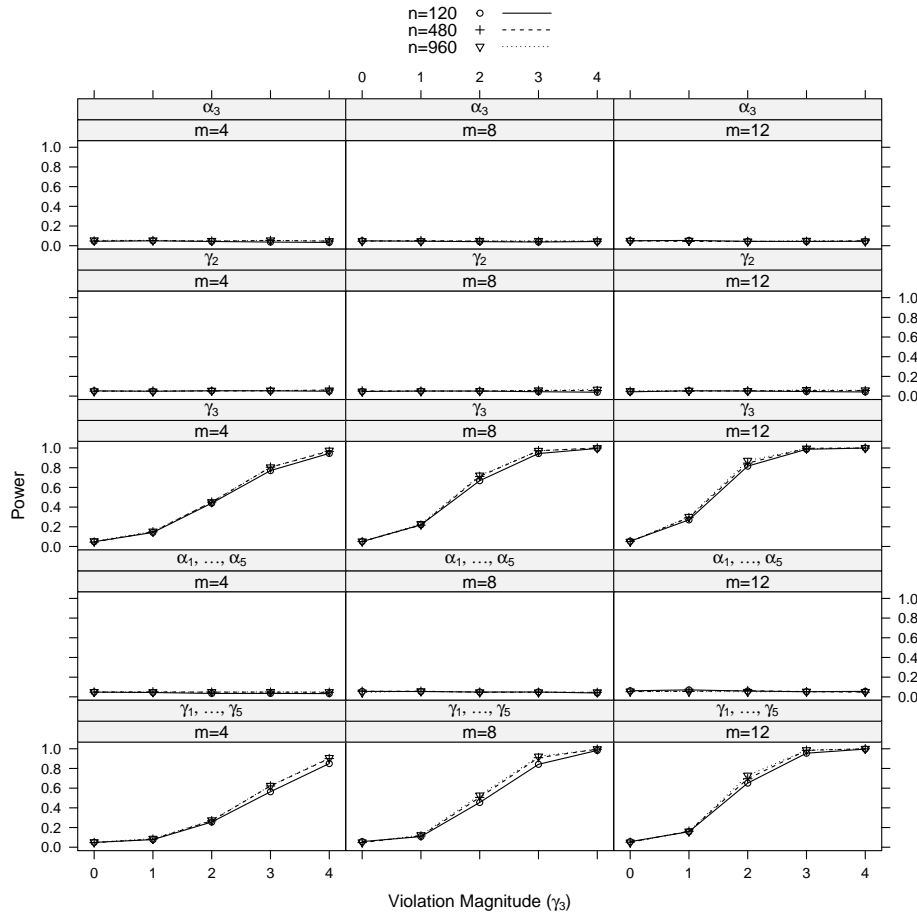


Figure 4: Simulation 1. Simulated power curves for sample sizes  $n = 120, 480$  and  $960$  of test statistic  $WDM_\rho$ , across three levels of the ordinal variable  $m$  and measurement invariance violations of 0–4 standard errors (scaled by  $\sqrt{n}$ ), estimated by PML two-parameter item response model. The parameter violating measurement invariance is  $\gamma_3$ . Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable  $m$ .

## 7. Simulation 2

In Simulation 1, the ability distributions were assumed to be the same for all persons. This ignored the fact that person hyperparameters (mean ability, variance of ability) could change across groups along with the item parameters. Changes in person hyperparameters do not count as measurement invariance violations, but ignoring these changes may lead us to incorrectly conclude an invariance violation Woods (2009); Stark, Chernyshenko, and Drasgow (2006); Wang and Yeh (2003); Fischer (1995a); Kopf, Zeileis, and Strobl (2015).

Formally, in a regular two-parameter model, we assume that the person parameters follow a standard normal distribution across all groups:  $\theta_i \sim N(0, 1)$ . There is the potential that the hyperdistribution is group specific, however, with  $\theta_i^* \sim N(\mu_{v_i}, \sigma_{v_i}^2)$ , where  $v_i$  is in  $1, \dots, m$ . If

the hyperparameters change from group to group, then our model can be written as:

$$\Phi^{-1}(p_{ij}) = \gamma_j + \alpha_j \theta_i^*, \quad (29)$$

$$= \gamma_j + \alpha_j (\mu_{v_i} + \sigma_{v_i} \theta_i), \quad (30)$$

$$= (\gamma_j + \alpha_j \mu_{v_i}) + \sigma_{v_i} \alpha_j \theta_i. \quad (31)$$

This shows that, when  $\sigma_{v_i}$  differs across values of  $v_i$ , it will look like there are measurement invariance violations in  $\alpha_j$  (for all  $j$ ). Similarly, when  $\mu_{v_i}$  differs across values of  $v_i$ , it will look like there are measurement invariance violations in  $\gamma_j$  (for all  $j$ ). Further, because  $\mu_{v_i}$  is no longer 0, changes in  $\alpha_j$  will also make it look like there are measurement invariance violations in the  $\gamma_j$  (through the term  $\alpha_j \mu_{v_i}$ ). Therefore, the proposed tests' good properties from Simulation 1 are lost when the person hyperparameters change across groups.

To avoid this problem, we should estimate the person hyperparameters  $\mu_{v_i}$  and  $\sigma_{v_i}^2$ , when there is uncertainty about person abilities. It is clear that estimation of these extra parameters will decrease the proposed tests' power. However, both the extent of decrease and the relative performance compared to traditional statistics is unclear. In this section, we conduct two simulations that address these issues.

## 7.1. Method

To examine the decrease in power when we estimate person hyperparameters with or without a "true" person hyperparameter change, we organize Simulation 2 into two subsections. In Simulation 2.1, the data-generation model is the same as Simulation 1, with abilities of students generated from  $\theta_i \sim N(0, 1)$  whereas, in Simulation 2.2, the abilities of students were manipulated. Specifically, abilities of students with  $V = 1, 2, 3$ , or 4 were generated from  $\theta_i \sim N(0, 1)$ , while the abilities of students with  $V = 5, 6, 7$ , or 8 were generated from  $\theta_i \sim N(-1, 2)$ .

The estimated model for both Simulations 2.1 and 2.2 is the multiple-group two-parameter model, which can be described as: free parameters for each level's  $\mu_{v_i}$  (with level 1 fixed to zero for identification),  $\sigma_{v_i}^2$  (with level 1 fixed to 1 for identification) and the five items' slope and intercept parameters (as in Simulation 1), with estimates again being obtained by PML.

Because the multiple-group two-parameter model has more parameters to be estimated (7 mean parameters  $\mu_{v_i}$  and 7 variance parameters  $\sigma_{v_i}^2$ ), the sample sizes were increased to  $n = 1200, 4800$ , and 9600. Measurement invariance violations still occurred in the same places (either  $\alpha_3$  or  $\gamma_3$ ), and the subsets of tested parameters were the same as in Simulation 1.

Power and type I error were examined across three sample sizes and 17 magnitudes of invariance violations (manipulated in the same way as Simulation 1). For each combination of sample size ( $n$ )  $\times$  violation magnitude ( $d$ ), 5000 data sets were generated and tested. In all conditions, we still maintained equal sample sizes in each level of  $V$ . We examined the statistics from Equations (26), (27) and (28).

## 7.2. Results

In the sections below, we first discuss results when the data-generation model had person hyperparameters that were the same across groups (Simulation 2.1). We then discuss results

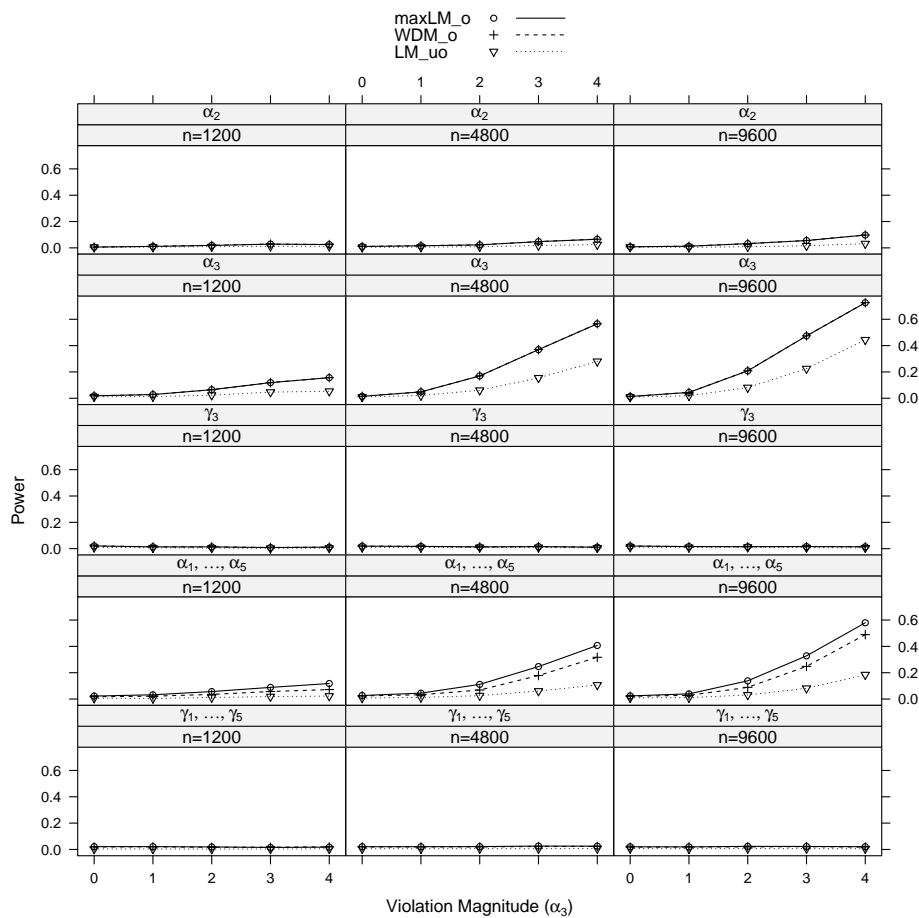


Figure 5: Simulation 2.1. Simulated power curves for  $\max LM_o$ ,  $WDM_o$ , and  $LM_{uo}$  across measurement invariance violations of 0–4 standard errors (scaled by  $\sqrt{n}$ ), estimated by PML (fitting multiple-group two-parameter model, without person abilities change in the generation model). The parameter violating measurement invariance is  $\alpha_3$ . The number of categories is  $m = 8$ . Panel labels denote the parameter(s) being tested and sample size.

when the data-generation model had person hyperparameters that differed across groups (Simulation 2.2).

### Simulation 2.1

Results are presented in Figures 5 and 6. Figure 5 demonstrates power curves as a function of violation magnitude in item 3’s slope parameter  $\alpha_3$ , with the parameters being tested changing across rows, the sample sizes  $n$  changing across columns, and lines reflecting different test statistics. Figure 6 demonstrates similar power curves when the violating parameter is item 3’s intercept parameter  $\gamma_3$ . In both figures, tests of item 2’s parameters are again representative of all invariant items.

From these two figures, one generally observes that the tests isolate the parameter violating measurement invariance in the multiple-group two-parameter model (across rows), and power

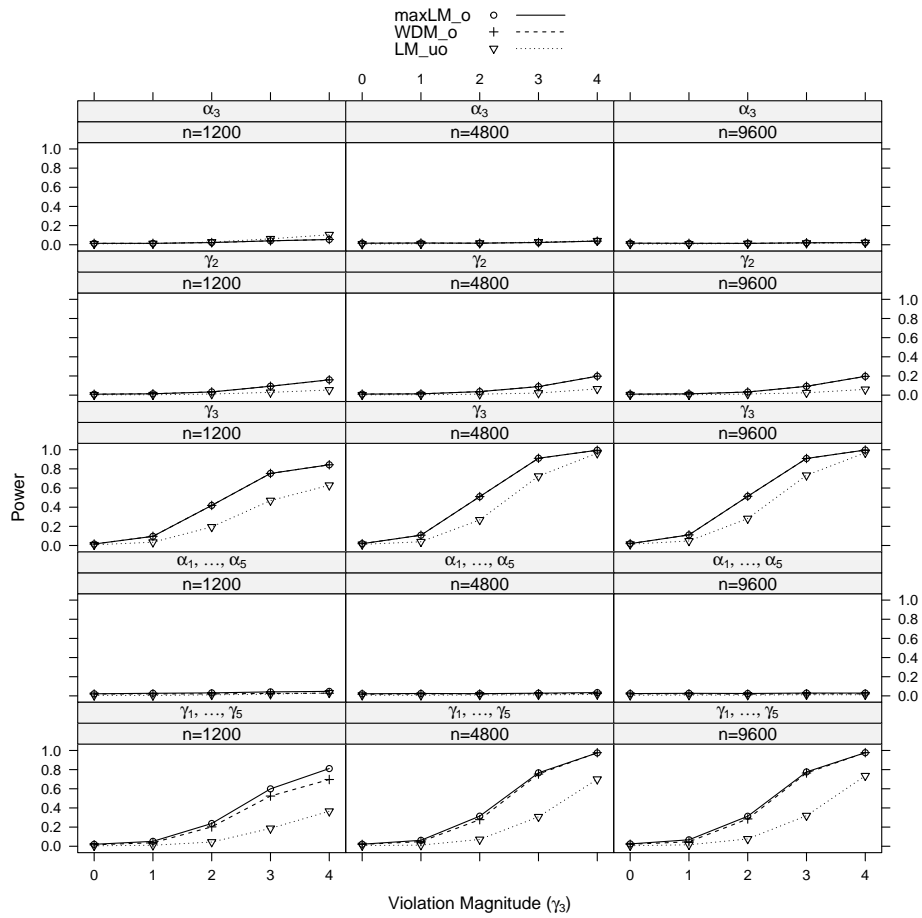


Figure 6: Simulation 2.1. Simulated power curves for  $\max LM_o$ ,  $WDM_o$ , and  $LM_{uo}$  across measurement invariance violations of 0–4 standard errors (scaled by  $\sqrt{n}$ ), estimated by PML (fitting multiple-group two-parameter model, without person abilities change in the generation model). The parameter violating measurement invariance is  $\gamma_3$ . The number of categories is  $m = 8$ . Panel labels denote the parameter(s) being tested and sample size.

increases with  $n$  (across columns). The impact of  $n$  is more substantial when the slope parameter, as opposed to the intercept parameter, violates invariance. We need sample size as large as 9600 to obtain power near .8 for detecting DIF in the slope parameter (with increasing violation magnitude), whereas there is no large difference across columns when the intercept parameter violates invariance.

Within each panel of Figures 5 and 6, the three lines reflect the three test statistics. It is seen that the two ordinal statistics still exhibit similar results, with  $\max LM_{uo}$  demonstrating lower power across all situations. Therefore, the sensitivity of the ordinal statistics is preserved in the multiple-group two-parameter model.

Comparing Figure 5 and Figure 6 in general, we can see the tests still have somewhat higher power to detect measurement invariance violations in the intercept parameter as opposed to the slope parameter. Moreover, power is lower when we test the full set of slope (or intercept) parameters, as opposed to only the problematic parameter.



*Simulation 2.2*

Results are presented in supplementary material, with the same figure and panel arrangements as Simulation 2.1. They demonstrate the same pattern as Simulation 2.1. We can observe that the power decrease is related to the number of parameters in the estimated model, regardless of the data-generation model.

In summary, we found that the proposed tests can attribute measurement invariance violations to the correct multiple-group model parameter when impact is exhibited. Although the multiple-group model requires a much larger sample size to obtain reasonable power, this type of model is necessary in practice when there is uncertainty about changes in person hyperparameters. Otherwise, there will be a serious “false alarm” as illustrated by Equations 29 to 31. The sample size issue can often be addressed, as IRT researchers often have thousands of respondents in their datasets. However, in other situations, we may wish to first test whether the hyperparameters vary across groups before examining the item parameters. If this test indicates that the hyperparameters do not vary, then we can constrain them to be equal and gain more power to detect DIF. If the test indicates that the hyperparameters do vary, then we can use the “location” information resulting from the tests to potentially reduce the number of hyperparameters in the model (i.e., by constraining similar groups’ parameters to be equal). This would again lead to increased power to detect DIF.

The procedure outlined in the previous paragraph has the potential to capitalize on chance, as we are relying on sequential statistical tests to modify the focal model. However, sequential statistical tests are commonly used in the DIF literature for, e.g., anchor item selection and item “purification”. Because the tests proposed here utilize a constrained model where all items serve as anchors, we have essentially replaced the sequential anchor item tests with sequential tests of the person hyperparameters. Further, to address concerns about capitalizing on chance, we can employ cross-validation methods. These strategies are demonstrated below in a practical example.

## 8. Application

We illustrate the tests’ application using 18 dichotomously scored mathematics items from the graduation examination developed by the Netherlands National Institute for Educational Measurement [Doolaard \(1999\)](#); [Fox \(2010\)](#).

### 8.1. Method

In the data set, 2156 eighth grade students completed the test, with a socioeconomic status (SES) variable also being measured on each student. The SES scores were based on four indicators, which were the education and occupation levels of both parents (if present). In this sample, there are 40 unique SES values ranging from  $-3.23$  to  $2.8$ , with higher values indicating higher SES. For the purposes of demonstration, we treat SES as a 6-category ordinal variable here and maintain equal sample sizes at each level.

The correlation between SES and mathematics achievement (sum of the 18 items) equals 0.49. Of course, this relationship could be explained in two different manners: either people of different SES exhibit different abilities, or the items are unfair to people of certain SES levels. We use the score-based tests to distinguish between these different explanations.

Following the strategy outlined at the end of the previous section, we start with a two-parameter item response model where the person hyperparameters  $\mu_1$  and  $\sigma_1^2$  (for level 1) are fixed to 0 and 1, while the hyperparameters in other levels are estimated but constrained to be equal, in the following referred to as the constrained hyperparameter model. This allows us to test whether the hyperparameters are equal across levels, and, if hyperparameters are not equal, it provides us with information about specific groups that are unequal. This information is used to build a model with relatively higher power to detect DIF and avoids “false alarm” by accounting for person hyperparameters. To address the potential problem of “capitalizing on chance” by adopting this strategy, overall model fit and cross validation are examined.

## 8.2. Results

We describe the results in three sections, one for the initial examination of fluctuations in the hyperparameters, one for examination of parameters based on the model with appropriate hyperparameters, and one for further support of our chosen model.

### *Testing the hyperparameters*

Results representing the statistics’ fluctuations across SES level based on the constrained hyperparameter model are shown in Figure 7. The first column displays the fluctuation process associated with  $LM_o$  for testing the 18 items’ slopes (first row), the 18 items’ intercepts (second row), the person mean parameters (third row), and the person variance parameters (fourth row). The second column displays the fluctuation process associated with  $WDM_o$  for the same sets of parameters. In other words, these panels show the values of Equations (26) and (27) for each SES level, with the dashed horizontal line being the 5% critical value. If the solid line crosses the critical value, then there is evidence that the corresponding parameter fluctuates across levels of SES. Because the final level’s statistics always equal zero (see Equation (11)), the final level (level 6 here) is not displayed.

It is observed that the person mean parameter (third row) fluctuates across all levels, while the person variance parameter (fourth row) fluctuates between the middle levels and level 5 (note that person hyperparameter change is not DIF). As shown in Simulation 2, this can cause the slope (first row) and intercept (second row) parameters to exhibit DIF regardless of whether they actually exhibit DIF. Therefore, we need to examine a second model where person hyperparameters are free across specific levels of SES. Based on the statistics’ fluctuation processes, the second model should estimate a separate person mean parameter for each SES level and a separate person variance parameter for the middle levels (levels 2–4) and for the extreme levels (at and after level 5). The test results involving this partially-free hyperparameter model is described in the next section.

### *Testing the partially-free hyperparameter model*

In estimating a separate  $\mu_{v_i}$  for each of the six SES groups (with first level being fixed to 0 for identification) and two separate  $\sigma^2$ s for the middle level and extreme levels, we obtain the results shown in Figure 8. The panel arrangements are the same as Figure 7.

Figure 8 implies that no sets of item parameters exhibit DIF, according to either statistic. This is the opposite result of what we found in the previous section, and it is related to the

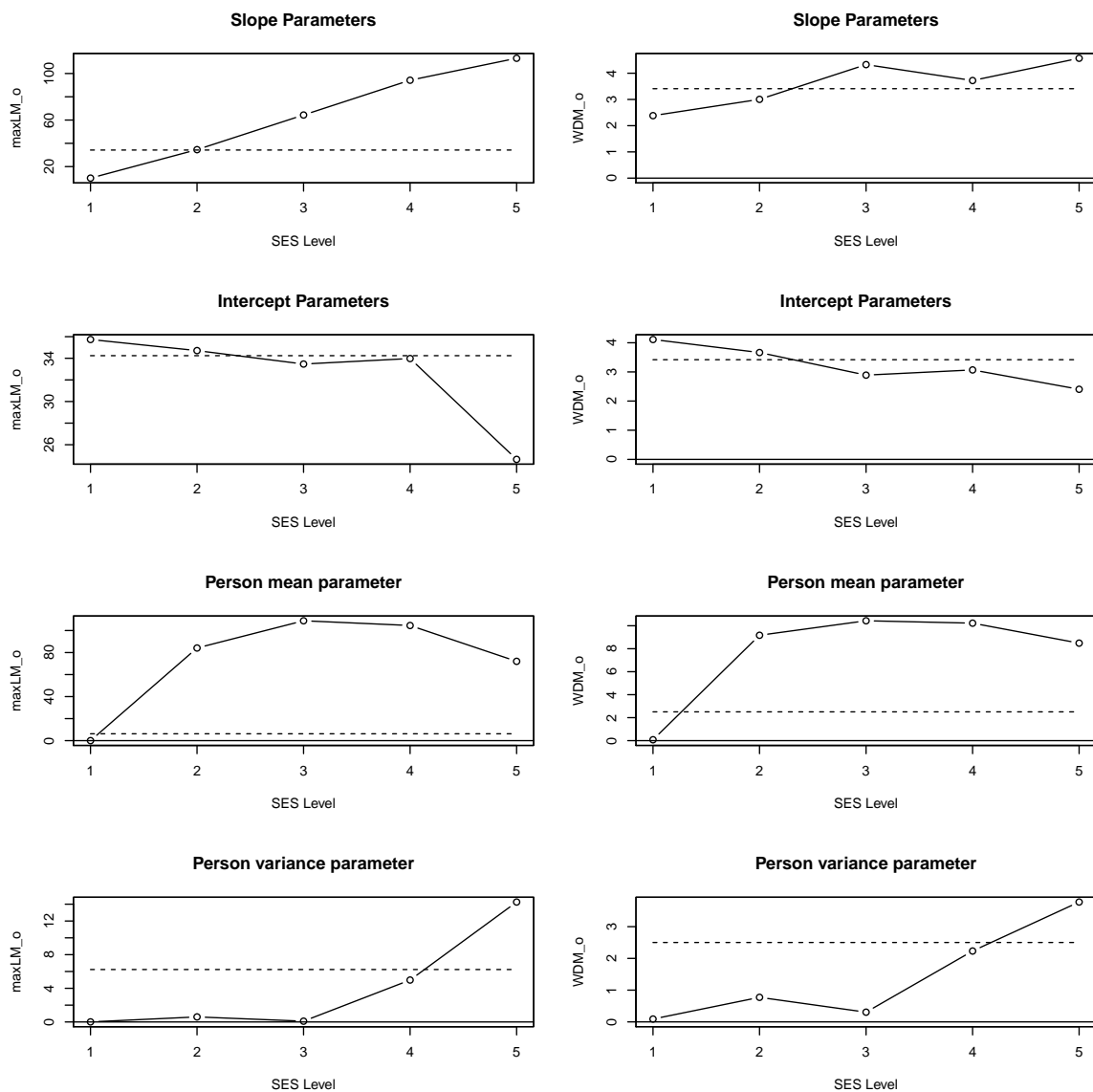


Figure 7: Empirical fluctuation processes of the max  $LM_o$  statistic (first column) and  $WDM_o$  (second column) for slope parameters (first row), intercept parameters (second row), person mean parameter (third row) and person variance parameter (fourth row), using constrained hyperparameter model.

findings from Simulation 2. Further, the estimated  $\mu_{v_i}$  increase monotonically with SES, with the lowest SES level having a fixed mean of 0, followed by 0.54, 1.01, 1.26, 1.58, and 2.25. Meanwhile,  $\sigma^2$  for the middle SES levels (levels 2–4) and extreme SES levels (level 5–6) are 1.14 and 1.37, with the lowest SES level having a fixed variance of 1.

*Overall model fit and cross validation*

As mentioned earlier, the fact that we sequentially studied person parameters and item pa-

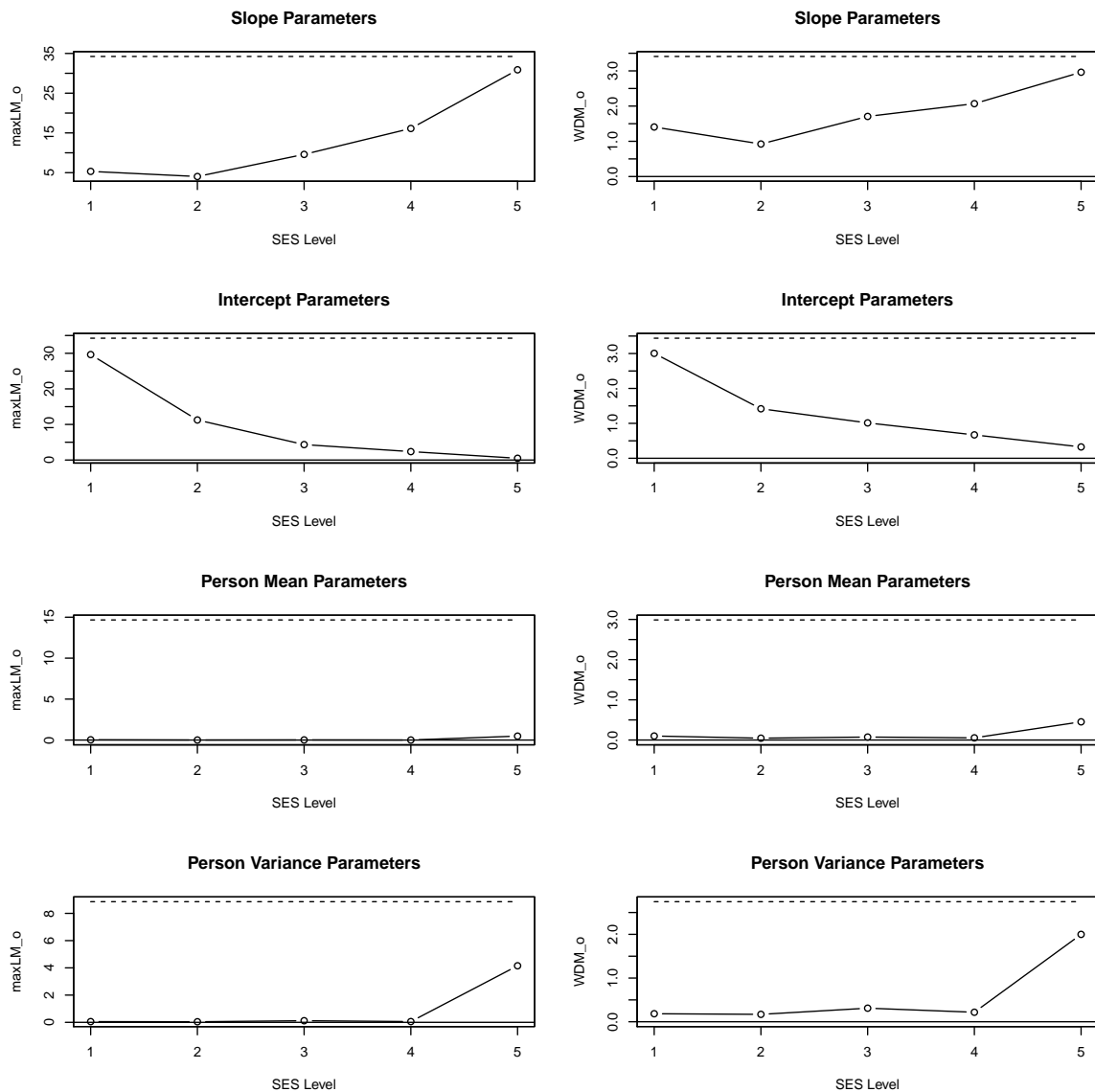


Figure 8: Empirical fluctuation processes of the max  $LM_o$  statistic (first column) and  $WDM_o$  (second column) for slope parameters (first row), intercept parameters (second row), person mean parameters (third row) and person variance parameters (fourth row), using partially-free hyperparameter model.

rameters is potentially problematic from the perspective of “capitalizing on chance”. This is because our tests of item parameters were based on a model that was influenced by the tests of person parameters. In this section, we do model comparisons and cross-validations to examine the extent to which our results were robust.

We start with general model comparisons. In addition to the two models examined above, we added a third model where each SES level has unique hyperparameters (similar to the model from Simulation 2). As mentioned before, this model will generally have lower power compared to the partially-free hyperparameter model, but it also avoids “false alarm” to the

	AIC <sub>PL</sub>	BIC <sub>PL</sub>		PLRT ( <i>p</i> -value)
constrained	733831.17	736465.92	constrained vs. partially-free	346.73 (0.00)
partially-free	718893.34	723281.20	partially-free vs. fully-free	3.84 (0.27)
fully-free	718907.81	723408.24	fully-free vs. constrained	380.80 (0.00)
Training replication (model ordering)	86	100	Training replication (partially-free vs. fully-free)	90
Test replication (model ordering)	97	97	Test replication (partially-free vs. fully-free)	100

Table 1: Model fit, comparison, and replication statistics from the Application. The labels “constrained”, “partially-free” and “fully-free” represent the constrained hyperparameter model, partially-free hyperparameter model and fully-free hyperparameter model, respectively. The bottom half of the table contains the percentage of 100 cross-validations for which the results replicated the original results (in order of model preference and the PLRT of the top two models).

greatest extent. This third model is referred as the fully-free hyperparameter model in the following. Both the AIC and BIC statistics (arising from PML, denoted as AIC<sub>PL</sub> and BIC<sub>PL</sub> below) preferred the partially-free hyperparameter model to the fully-free hyperparameter model, as well as the constrained hyperparameter model. Model statistics are given in the top half of Table 1. In addition, the pairwise likelihood ratio test (PLRT) preferred the partially-free hyperparameter model to the other two models. Specifically, the fit of the fully-free hyperparameter model is not significantly better than the partially-free hyperparameter model (3.84,  $p = 0.27$ ) and the partially-free model is significantly better than that of the constrained hyperparameter model (346.73,  $p = 0.00$ ). Additionally, the model with all free hyperparameters is preferred to the constrained hyperparameter model (380.80,  $p = 0.00$ ). Thus, the preference order of these three models is: partially-free hyperparameter model, then fully-free hyperparameter model, followed by the constrained hyperparameter model.

In order to confirm the generality of the above model assessment, we conducted a cross-validation whereby half of the original dataset was randomly allocated to the training set, with the remaining half being allocated to the test set. This random allocation is replicated 100 times, resulting 100 training data sets and corresponding 100 test data sets. For each of 100 training data sets, we fitted the three models and compared them to one another. We then computed analogous model fit statistics for the test data, holding the model parameter values at the estimates from the training data. This allows us to examine the extent to which the fitted models continue to be preferred in new data. These results are displayed in the bottom half of Table 1, where the numbers represent percentages of the 100 datasets that replicated the original results.

The bottom half of the table shows that, in the training data sets, the original model preference ordering was replicated 86 times out of 100 based on AIC<sub>PL</sub> and 100 times out of 100 based on BIC<sub>PL</sub>. This result is also supported by pairwise log-likelihood ratio test when comparing the top two models, whereby 90 out of 100 statistics preferred the partially-free hyperparameter model at  $\alpha = .05$ . Similarly, in the test data sets, the same model preference ordering was replicated 97 times out of 100 based on AIC<sub>PL</sub>, as well as 97 times out of 100 based on BIC<sub>PL</sub>.

Finally, applying score-based tests to the partially-free hyperparameter model for the training data, 83% (using the statistic  $WDM_o$ ) and 99% (using the statistic  $LM_o$ ) datasets exhibited no DIF. Taken together, these analyses illustrate that our original results (model preferences and DIF results) remain similar across the 100 resamples.

In summary, we found that the positive correlation between SES and math achievement is due to the fact that students' ability means and variances increase with SES. All parameters appear to fulfill the measurement invariance assumption after we take account of changes in person ability at corresponding SES level. The score-based tests allowed us to systematically study these issues without estimating an excessive number of models. If desired, we could also test each item's parameters individually (as opposed to the set of intercepts and the set of slopes) without fitting any new models. This illustrates the inherent flexibility of the tests.

## 9. General discussion

In this paper, we extended a recently proposed family of score-based tests to item response models, focusing on multiple-group two-parameter models. The tests' power levels are comparable to traditional statistics, and the tests can isolate specific parameters violating invariance so long as we account for changes in person ability across groups.

The test statistics examined here, along with estimation by PML, provides a more general and flexible framework to detect DIF in IRT research. Traditionally, we pre-define two groups of individuals and compare them via a multiple-group model. In using score-based tests, we do not need to pre-define the groups and can test many groups simultaneously. Additionally, person hyperparameters can be estimated conveniently in a multiple-group null model (that assumes measurement invariance holds) without re-fitting multiple alternative models as is required by the LRT or Wald test (see also Glas 1998). This can enhance our ability to detect DIF in large datasets with many groups.

In the sections below, we consider the tests' applications in related models and in complex scenarios.

### 9.1. Model extensions

The PML framework generally allows us to use the score-based tests in situations when the responses have multiple categories, where a graded response model Samejima (1969) or partial credit model Muraki (1992) may be used. These models become increasingly difficult to estimate when we have many groups and when items have many categories. In these situations, the score-based tests become increasingly attractive because they require estimation of only a null model (assuming that invariance holds).

Another extension involves use of multidimensional IRT models, especially because multidimensionality is one possible cause of DIF Millsap (2012). However, it is difficult to test this hypothesis due to the multidimensional integration involved. In employing the factor-analytic framework described here with PML, we can more easily estimate models with multiple dimensions. This can further help us study DIF in larger datasets.

Finally, moving beyond traditional IRT models, the tests proposed here can be applied to multilevel/mixed models where, e.g., students' responses may be nested in classes, schools, or states. Score-based tests only rely on the derivative of each individual's likelihood function

so that, as long as the individual derivative (analytic or approximation) can be specified, the tests can be applied. Scores for generalized linear mixed models will be more difficult to obtain than scores for linear mixed models, in the same way that scores for continuous-data factor analysis are easier to obtain than scores for IRT models.

## 9.2. Full structural equation modeling approach to linking/equating problem

In practice, we often need to transform person parameters so that ability estimates are equivalent across different scales. This is called equating (see [Kolen and Brennan 2004](#), for a review). For example, we may need to equate test takers' abilities across multiple versions of the SAT.

The existence of DIF complicates equating. Suppose that Form A of the SAT exhibits DIF with respect to country/grade/age, but Form B does not exhibit DIF. We must then decide whether we should equate each level of  $V$  separately, as opposed to equating simultaneously across the whole sample. [Dorans \(2004\)](#) dealt with this question by introducing new statistics that utilized the test characteristic curve. Alternatively, we can frame the question in a full structural equation model (SEM) and employ the score-based test to examine the corresponding coefficients' stability against  $V$ . In this way, no new statistics need to be introduced.

## 9.3. Multiple violating slope parameters

In this paper, we studied the tests' applications to two-parameter and multiple-group two-parameter models when only one parameter violated invariance. When there are multiple violating parameters, [Bechger and Maris \(2015\)](#) point out that both the null and alternative hypotheses of a score-based test can be incorrect. For example, if we test a single item intercept parameter, then the null hypothesis would involve all intercept parameters being equal across groups and the alternative hypothesis would involve the focal intercept parameter being unequal across groups (with the remaining intercepts being equal). If a non-focal intercept parameter is unequal across groups, however, then both hypotheses are incorrect.

To address this issue, we can employ recursive tests related to item purification. This could proceed as follows (see [Glas 1998](#), for a related approach): (1) fit the null model with person hyperparameters, (2) test for DIF in each item parameter, (3) free the parameter with the largest statistic and refit the model with person hyperparameters, (4) repeat steps (2)–(3) until there is no further DIF detected. This procedure is similar to the LRT algorithm described by [Magis \*et al.\* \(2010\)](#), which is implemented in R packages `mirt` [Chalmers \(2012\)](#) and `diffR` [Magis, Beland, and Raiche \(2015\)](#). The score-based tests are advantageous here because no anchor items are needed (see [Woods 2009](#), for a review of procedures involving anchor items). This is because we only need to estimate the null model, where all parameters are already assumed to be invariant across groups. However, the sensitivity to the order of purification described by, e.g., [Magis and Facon \(2013\)](#) and [Bechger and Maris \(2015\)](#) cannot be avoided under this approach.

As an alternative to the item purification approach, [Bechger and Maris \(2015\)](#) make the insightful point that, in a Rasch framework, pairwise differences between item parameters are preserved across the set of possible identification constraints. Thus, they conceptualize *differential item pair functioning* as a property of item pairs, whereby differences between item parameters may vary across groups (as opposed to individual item parameters varying

across groups). This proposal leads to Wald tests of differences between item parameters, where the test results are the same regardless of choice of identification constraint. A potential difficulty here is that, in the marginal and pairwise ML frameworks considered in this paper, we typically want hyperparameters to be free across groups, and estimation of these hyperparameters requires more parameter constraints than would typically be employed in the Wald test framework. It appears difficult to address all these issues without an iterative procedure.

Nonetheless, we might make some progress through consideration of alternative parameter constraints. That is, instead of constraining one group’s mean and variance hyperparameters to 0 and 1, respectively, we may employ “sum” constraints that allow us to freely estimate more parameters. For example, Verhagen, Levy, Millsap, and Fox (2016) constrained the sum of all intercept parameters to be zero (in a Rasch-type model) to avoid the need for defining anchor items or assuming group ability (i.e. fixing one group ability parameter). These constraints can be extended to the slopes of a two-parameter model, requiring that the squared slope parameters sum to 1. Further work may consider the combination of these types of parameter constraints with both score-based tests and differential item pair functioning.

#### 9.4. Summary

In this paper, we generalized the score-based tests to IRT models estimated by MML and PML. This extension has advantages over traditional DIF detection methods in locating the violating parameter without pre-specifying grouping information and in accounting for the ordinal information of the auxiliary variable  $V$ . Besides, implementation of these tests is simpler, requiring only estimation of a null model that assumes measurement invariance. Applied researchers in psychology and education could use these tests to conveniently examine measurement invariance in their own data sets.

## Computational details

All results were obtained using the R system for statistical computing R Core Team (2017), version 3.2.3, employing the add-on package `lavaan` 0.5-23.1097 Rosseel (2012) for fitting of the factor analysis models and `strucchange` 1.5-1 Zeileis, Leisch, Hornik, and Kleiber (2002); Zeileis (2006) for evaluating the parameter instability tests. R and both packages are freely available under the General Public License from the Comprehensive R Archive Network at <https://CRAN.R-project.org/>. R code for replication of our results is available in the supplementary materials and at <http://semtools.R-Forge.R-project.org/>.

## References

- Andrews DWK (1993). “Tests for Parameter Instability and Structural Change with Unknown Change Point.” *Econometrica*, **61**, 821–856. doi:10.2307/2951764.
- Bechger TM, Maris G (2015). “A Statistical Test for Differential Item Pair Functioning.” *Psychometrika*, **80**(2), 317–340. doi:10.1007/s11336-014-9408-y.



- Bock RD, Schilling S (1997). “High-Dimensional Full-Information Item Factor Analysis.” In M Berkane (ed.), *Latent Variable Modeling and Applications to Causality*, pp. 163–176. Springer-Verlag, New York. doi:10.1007/978-1-4612-1842-5\_8.
- Chalmers RP (2012). “**mirt**: A Multidimensional Item Response Theory Package for the R Environment.” *Journal of Statistical Software*, **48**(6), 1–29. doi:10.18637/jss.v048.i06.
- De Ayala RJ (2009). *The Theory and Practice of Item Response Theory*. Guilford Press, New York.
- Doolaard S (1999). “Schools in Change or Schools in Chains.” *Unpublished doctoral dissertation, University of Twente, The Netherlands*. doi:10.1016/j.gloplacha.2013.03.013.
- Dorans NJ (2004). “Using Subpopulation Invariance to Assess Test Score Equity.” *Journal of Educational Measurement*, **41**(1), 43–68. doi:10.1111/j.1745-3984.2004.tb01158.x.
- Fischer GH (1995a). “Derivations of the Rasch Model.” In GH Fischer, IW Molenaar (eds.), *Rasch Models*, pp. 15–38. Springer-Verlag, New York. doi:10.1007/978-1-4612-4230-7\_2.
- Fischer GH (1995b). “Some Neglected Problems in IRT.” *Psychometrika*, **60**(4), 459–487. doi:10.1007/bf02294324.
- Fischer GH, Molenaar IW (2012). *Rasch Models: Foundations, Recent Developments, and Applications*. Springer-Verlag. doi:10.1007/978-1-4612-4230-7.
- Fox JP (2010). *Bayesian Item Response Modeling: Theory and Applications*. Springer-Verlag, New York. doi:10.1007/978-1-4419-0742-4.
- Glas CAW (1998). “Detection of Differential Item Functioning Using Lagrange Multiplier Tests.” *Statistica Sinica*, **8**(3), 647–667. doi:10.1007/978-1-4613-0169-1\_8.
- Glas CAW (1999). “Modification Indices for the 2-PL and the Nominal Response Model.” *Psychometrika*, **64**(3), 273–294. doi:10.1007/bf02294296.
- Glas CAW (2009). “Item Parameter Estimation and Item Fit Analysis.” In WJ Van der Linden, CAW Glas (eds.), *Elements of Adaptive Testing*, pp. 269–288. Springer-Verlag, New York. doi:10.1007/978-0-387-85461-8\_14.
- Glas CAW (2010). “Testing Fit to IRT Models for Polytomously Scored Items.” In ML Nering, R Ostini (eds.), *Handbook of Polytomous Item Response Theory Models*, pp. 185–210. Routledge, New York. doi:10.4324/9780203861264.ch8.
- Glas CAW (2015). “Item Response Theory Models in Behavioral Social Science: Assessment of Fit.” *Wiley StatsRef: Statistics Reference Online*. doi:10.1002/9781118445112.stat06436.pub2.
- Glas CAW, Falcón JCS (2003). “A Comparison of Item-Fit Statistics for the Three-Parameter Logistic Model.” *Applied Psychological Measurement*, **27**(2), 87–106. doi:10.1177/0146621602250530.

- Glas CAW, Jehangir K (2014). “Modeling Country-Specific Differential Item Functioning.” In L Rutkowski, M von Davier, D Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, pp. 97–115. Chapman and Hall/CRC, Boca Raton, FL. doi:10.1111/jedm.12095.
- Glas CAW, Van der Linden WJ (2010). “Marginal Likelihood Inference for a Model for Item Responses and Response times.” *British Journal of Mathematical and Statistical Psychology*, **63**(3), 603–626. doi:10.1348/000711009x481360.
- Hjort NL, Koning A (2002). “Tests for Constancy of Model Parameters over Time.” *Non-parametric Statistics*, **14**, 113–132. doi:10.1080/10485250211394.
- Holland PW, Thayer DT (1988). “Differential Item Performance and the Mantel-Haenszel Procedure.” In H Wainer, HI Braun (eds.), *Test Validity*, pp. 129–145. Routledge, Hillsdale.
- Katsikatsou M, Moustaki I (2016). “Pairwise Likelihood Ratio Tests and Model Selection Criteria for Structural Equation Models with Ordinal Variables.” *Psychometrika*, **81**(4), 1046–1068. doi:10.1007/s11336-016-9523-z.
- Katsikatsou M, Moustaki I, Yang-Wallentin F, Jöreskog KG (2012). “Pairwise Likelihood Estimation for Factor Analysis Models with Ordinal Data.” *Computational Statistics & Data Analysis*, **56**(12), 4243–4258. doi:10.1016/j.csda.2012.04.010.
- Kolen MJ, Brennan RL (2004). *Test Equating, Scaling, and Linking*. Springer-Verlag, New York. doi:10.1007/978-1-4757-4310-4\_10.
- Kopf J, Zeileis A, Strobl C (2015). “Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches.” *Educational and Psychological Measurement*, **75**(1), 22–56. doi:10.1177/0013164414529792.
- Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems*. Routledge, New York. doi:10.4324/9780203056615.
- Magis D, Beland S, Raiche G (2015). *difR: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF)*. doi:10.3758/brm.42.3.847. R package version 4.6.
- Magis D, Béland S, Tuerlinckx F, De Boeck P (2010). “A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning.” *Behavior Research Methods*, **42**(3), 847–862. doi:10.3758/brm.42.3.847.
- Magis D, Facon B (2013). “Item Purification Does Not Always Improve DIF Detection: A Counterexample with Angoff’s Delta Plot.” *Educational and Psychological Measurement*, **73**(2), 293–311. doi:10.1177/0013164412451903.
- Mellenbergh GJ (1989). “Item Bias and Item Response Theory.” *International Journal of Educational Research*, **13**, 127–143. doi:10.1016/0883-0355(89)90002-5.
- Merkle EC, Fan J, Zeileis A (2014). “Testing for Measurement Invariance with Respect to an Ordinal Variable.” *Psychometrika*, **79**, 569–584. doi:10.1007/s11336-013-9376-7.
- Merkle EC, Zeileis A (2013). “Tests of Measurement Invariance without Subgroups: A Generalization of Classical Methods.” *Psychometrika*, **78**, 59–82. doi:10.1007/s11336-012-9302-4.

- Millsap RE (2005). “Four Unresolved Problems in Studies of Factorial Invariance.” In A Maydeu-Olivares, JJ McArdle (eds.), *Contemporary Psychometrics*, pp. 153–171. Lawrence Erlbaum Associates, Mahwah.
- Millsap RE (2012). *Statistical Approaches to Measurement Invariance*. Routledge, New York. doi:10.4324/9780203821961.
- Millsap RE, Everson HT (1993). “Methodology Review: Statistical Approaches for Assessing Measurement Bias.” *Applied Psychological Measurement*, **17**(4), 297–334. doi:10.1177/014662169301700401.
- Muraki E (1992). “A Generalized Partial Credit Model: Application of an EM Algorithm.” *Applied Psychological Measurement*, **16**, 159–176. doi:10.1177/014662169201600206.
- Osterlind SJ, Everson HT (2009). *Differential Item Functioning*, volume 161. Sage, Thousand Oaks, CA. doi:10.4135/9781412993913.
- Raju NS (1988). “The Area between Two Item Characteristic Curves.” *Psychometrika*, **53**(4), 495–502. doi:10.1007/bf02294403.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rosseel Y (2012). “lavaan: An R Package for Structural Equation Modeling.” *Journal of Statistical Software*, **48**(2), 1–36. doi:10.18637/jss.v048.i02.
- Samejima F (1969). “Estimation of Latent Ability Using a Response Pattern of Graded Scores.” *Psychometrika Monograph Supplement*. doi:10.1007/bf03372160.
- Satorra A (1989). “Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach.” *Psychometrika*, **54**, 131–151. doi:10.1007/bf02294453.
- Schilling S, Bock RD (2005). “High-Dimensional Maximum Marginal Likelihood Item Factor Analysis by Adaptive Quadrature.” *Psychometrika*, **70**(3), 533–555. doi:10.1007/s11336-003-1141-x.
- Stark S, Chernyshenko OS, Drasgow F (2006). “Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy.” *Journal of Applied Psychology*, **91**, 1292–1306. doi:10.1037/0021-9010.91.6.1292.
- Strobl C, Kopf J, Zeileis A (2015). “Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model.” *Psychometrika*, **80**, 289–316. doi:10.1007/s11336-013-9388-3.
- Swaminathan H, Rogers HJ (1990). “Detecting Differential Item Functioning Using Logistic Regression Procedures.” *Journal of Educational Measurement*, **27**(4), 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x.
- Takane Y, De Leeuw J (1987). “On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables.” *Psychometrika*, **52**, 393–408. doi:10.1007/bf02294363.

- Thissen D (1982). “Marginal Maximum Likelihood Estimation for the One-Parameter Logistic Model.” *Psychometrika*, **47**, 175–186. doi:10.1007/bf02296273.
- Thissen D, Steinberg L, Wainer H (1988). “Use of Item Response Theory in the Study of Group Differences in Trace Lines.” In H Wainer, HI Braun (eds.), *Test Validity*, pp. 147–172. Lawrence Erlbaum Associates, Hillsdale. doi:10.2307/1164765.
- Tutz G, Schauberger G (2015). “A Penalty Approach to Differential Item Functioning in Rasch Models.” *Psychometrika*, **80**(1), 21–43. doi:10.1007/s11336-013-9377-6.
- Van den Noortgate W, De Boeck P (2005). “Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models.” *Journal of Educational and Behavioral Statistics*, **30**(4), 443–464. doi:10.3102/10769986030004443.
- Verhagen J, Levy R, Millsap RE, Fox JP (2016). “Evaluating Evidence for Invariant Items: A Bayes Factor Applied to Testing Measurement Invariance in IRT Models.” *Journal of Mathematical Psychology*, **72**, 171–182. doi:10.1016/j.jmp.2015.06.005.
- Wang T, Merkle E, Zeileis A (2014). “Score-Based Tests of Measurement Invariance: Use in Practice.” *Frontiers in Psychology*, **5**(438), 1–11. doi:10.3389/fpsyg.2014.00438.
- Wang WC, Yeh YL (2003). “Effects of Anchor Item Methods on Differential Item Functioning Detection with the Likelihood Ratio Test.” *Applied Psychological Measurement*, **27**(6), 479–498. doi:10.1177/0146621603259902.
- Woods CM (2009). “Empirical Selection of Anchors for Tests of Differential Item Functioning.” *Applied Psychological Measurement*, **33**(1), 42–57. doi:10.1177/0146621607314044.
- Zeileis A (2006). “Implementing a Class of Structural Change Tests: An Econometric Computing Approach.” *Computational Statistics & Data Analysis*, **50**(11), 2987–3008. doi:10.1016/j.csda.2005.07.001.
- Zeileis A, Hornik K (2007). “Generalized M-Fluctuation Tests for Parameter Instability.” *Statistica Neerlandica*, **61**, 488–508. doi:10.1111/j.1467-9574.2007.00371.x.
- Zeileis A, Leisch F, Hornik K, Kleiber C (2002). “**strucchange**: An R Package for Testing Structural Change in Linear Regression Models.” *Journal of Statistical Software*, **7**(2), 1–38. doi:10.18637/jss.v007.i02.

### Affiliation:

Ting Wang  
 Department of Psychological Sciences  
 University of Missouri  
 Columbia, MO 65211, United States of America  
 E-mail: [twb8d@mail.missouri.edu](mailto:twb8d@mail.missouri.edu)

Carolin Strobl  
Department of Psychology  
University of Zurich  
Zurich, Switzerland  
E-mail: [c.strobl@psychologie.uzh.ch](mailto:c.strobl@psychologie.uzh.ch)

Achim Zeileis  
Department of Statistics  
Faculty of Economics and Statistics  
Universität Innsbruck  
Universitätsstr. 15  
6020 Innsbruck, Austria  
E-mail: [Achim.Zeileis@R-project.org](mailto:Achim.Zeileis@R-project.org)  
URL: <https://eeecon.uibk.ac.at/~zeileis/>

Edgar C. Merkle  
Department of Psychological Sciences  
University of Missouri  
Columbia, MO 65211, United States of America  
E-mail: [merklee@missouri.edu](mailto:merklee@missouri.edu)