Score-based tests of differential item functioning in the two-parameter model

Ting Wang, Carolin Strobl, Achim Zeileis,
Edgar C. Merkle

Working Papers in Economics and Statistics

2016-05

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
Research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

Score-Based Tests of Differential Item Functioning in the Two-Parameter Model

Ting Wang
University of Missouri

Carolin Strobl
University of Zurich

Achim Zeileis
Universität Innsbruck

Edgar C. Merkle
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 people. 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 estimated via 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: measurement invariance, item response theory, factor analysis, 2PL model, 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, e.g., 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):

$$f(\boldsymbol{y}_i | v_i, \boldsymbol{\eta}_i) = f(\boldsymbol{y}_i | \boldsymbol{\eta}_i), \tag{1}$$

where $\boldsymbol{y}_i$ is a vector of observed variables for individual i, $\boldsymbol{\eta}_i$ 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) $\boldsymbol{\eta}_i$ and response $\boldsymbol{y}_i$ (binary or ordinal) holds regardless of the value of V. Many previous procedures have been proposed to assess measurement invariance/DIF in IRT models (e.g., Lord 1980; Holland and Thayer 1988; Thissen, Steinberg, and Wainer 1988; Swaminathan and Rogers 1990; Rijmen, Tuerlinckx, Boeck, and Kuppens 2003; den Noortgate and Boeck 2005; Magis, Béland, Tuerlinckx, and Boeck 2010; Magis and Facon 2012), and 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).

A novel family of *score-based* tests has been proposed recently to address those "locating" issues in factor models for continuous data (Merkle and Zeileis 2013; Merkle, Fan, and Zeileis 2014; Wang, Merkle, and Zeileis 2014). 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.

In this paper, we extend the tests to more general IRT models in a unified way and focus on identifying violating item parameters without pre-specifying reference and focal groups. We first describe the two-parameter IRT model and its relationship to factor analysis, along with the score-based tests' application to IRT. 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 ($i \in 1, \ldots, n$) and j represents items ($j \in 1, \ldots, p$). 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

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(p_{ij}), && (2)\\ \text{logit}(p_{ij}) &= \alpha_j \eta_i + \gamma_j, && (3)\\ \eta_i &\sim N(\mu_i, \sigma_i^2), && (4) \end{aligned}$$

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 $\text{logit}(p_{ij}) = \log(\frac{p_{ij}}{1-p_{ij}})$, which is a linear function of the person's ability η_i and the item parameters γ_j and α_j. The alternative parameterization, $\alpha_j(\eta_i - \gamma_j)$, could also be used here. Finally, person ability η_i is described by hyperparameters μ_i and σ_i^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 $\Phi^{-1}()$ (the probit link function). In this case, Equation (3) could be written as $p_{ij} = \Phi(\alpha_j \eta_i + \gamma_j)$.

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

$$\boldsymbol{y}_i^\star = \boldsymbol{\Lambda}\eta_i + \boldsymbol{\epsilon}, \tag{5}$$

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

$$y_{ij} = \begin{cases} 1 & y_{ij}^\star \geq \tau_j \\ 0 & y_{ij}^\star < \tau_j. \end{cases} \tag{6}$$

Therefore, we can see that λ_j is similar to α_j in Equation (3); they are both attached to the ability variable η_i. The error term $\boldsymbol{\epsilon}$ is related to the probit link function that could be used in Equation (3). Finally, the threshold τ_j corresponds to γ_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 η_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; Ayala 2009), marginal maximum likelihood (MML) (Thissen 1982) and pairwise maximum likelihood (Katsikatsou, Moustaki, Yang-Wallentin, and Jöreskog 2012). We briefly describe each method below.

3. Estimation

3.1. CML

CML uses each person's sum score as a sufficient statistic for the person parameters. This allows us to condition on the sum score and avoid estimation of the person parameters. However, this property only holds for IRT models with $\alpha_j = 1$, such as the Rasch model (one-parameter model) and partial credit model. Since we aim to employ the two-parameter IRT model in this paper, CML cannot be used.

3.2. MML

MML extends to the two-parameter IRT model by integrating out person parameters. Specifically,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}_i) = \log \int_{-\infty}^{\infty} f(\boldsymbol{y}_i | \eta_i, \boldsymbol{\theta}) g(\eta_i) d\eta_i, \tag{7}$$

where $\ell(\boldsymbol{\theta}; \boldsymbol{y}_i)$ denotes the log-likelihood function for individual i under the model. The vector $\boldsymbol{\theta}$ contains the item parameters (throughout the paper, "parameter" refers to an item parameter if not otherwise specified), $g(\eta_i)$ is a normal distribution, and $f(\boldsymbol{y}_i|\eta_i, \boldsymbol{\theta})$ is binomial for an IRT model with binary data. The integral in Equation (7) has no analytical solution and is typically solved by quadrature. This makes it difficult to manipulate the likelihood functions and its derivatives.

3.3. PML

If we employ the factor analysis version of the model, the difficult integration occurs in a different place. Specifically, the log-likelihood function of individual i's observed data $\boldsymbol{y}_i$, given the parameter vector $\boldsymbol{\theta}$ (including λ, τ), is the integrals over $\boldsymbol{y}_i^\star$:

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}_i) \quad = \quad \log \int_{\boldsymbol{\tau}} f(\boldsymbol{y}_i^\star | \boldsymbol{\theta}) d\boldsymbol{y}_i^\star \tag{8}$$

where $\boldsymbol{y}_i^\star$ is described as Equation (5), the distribution of $\boldsymbol{y}_i^\star$ with η_i marginalized out is denoted as $f(\boldsymbol{y}_i^\star|\boldsymbol{\theta})$ (p dimensional), which can be considered as a multivariate normal distribution following $N(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})$. The integration of the p-dimensional multivariate normal distribution over support $\boldsymbol{\tau}$ is the difficult part, which does not have a closed form. To deal with this problem, commercial software generally relies on three-stage estimation methods (details see Muthén 1984; Jöreskog 1990).

Jöreskog and Moustaki (2001) proposed that the likelihood function above can be changed to:

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}_i) = \left\{ \sum_{j<k} \ell(\boldsymbol{\theta}; (y_{ij}, y_{ik})) - ap \sum_{j} \ell(\boldsymbol{\theta}; y_{ij}) \right\} \tag{9}$$

where $\sum_{j<k} \ell(\boldsymbol{\theta}; (y_{ij}, y_{ik}))$ is the log-likelihood associated with all pairs of items, which is a series of 2-way contingency tables. The parameter a is a constant to be chosen for optimal efficiency, commonly being set to 0 (Katsikatsou *et al.* 2012).

Thus, Equation (9) is reduced to the composite pairwise log-likelihood $p\ell(\boldsymbol{\theta}; \boldsymbol{y}_i)$ (Katsikatsou *et al.* 2012), which can be expressed as

$$\begin{aligned} p\ell(\boldsymbol{\theta}; \boldsymbol{y}_i) &= \left\{ \sum_{j<k} \ell(\boldsymbol{\theta}; (y_{ij}, y_{ik})) \right\}, & (10) \\ &= \left\{ \sum_{j<k} \left(\sum_{c_j=1}^{2} \sum_{c_k=1}^{2} \log \pi_{y_{ij} y_{ik}}^{(c_j c_k)}(\boldsymbol{\theta}) \right) \right\}, & (11) \end{aligned}$$

where $\pi_{y_{ij} y_{ik}}^{(c_j c_k)}(\boldsymbol{\theta})$ 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. Category $1, 2$ represents response of "0", "1" respectively in Equation (6).

For the ease and generalization of notation, we change Equation (6) to the following form:

$$y_{ij} = c_j \Longleftrightarrow \tau_j^{(c_j - 1)} < y_{ij}^\star \leq \tau_j^{(c_j)}, \tag{12}$$

where $-\infty = \tau_j^{(0)} < \tau_j^{(1)} < \tau_j^{(2)} = \infty$. We can see for binary data, only $\tau_j^{(1)}$ needs to be estimated, which is commonly referred as τ_j in Equation (6).

Thus $\pi_{y_{ij}y_{ik}}^{(c_j c_k)}(\boldsymbol{\theta})$ can be expressed explicitly in the following form:

$$\begin{aligned}\pi_{y_{ij}y_{ik}}^{(c_j c_k)}(\boldsymbol{\theta}) &= \pi(y_{ij} = c_j, y_{ik} = c_k; \boldsymbol{\theta}) \qquad (13)\\ &= \Big\{\Phi_2(\tau_j^{(c_j)}, \tau_k^{(c_k)}; \rho_{y_{ij}y_{ik}}) - \Phi_2(\tau_j^{(c_j)}, \tau_k^{(c_k-1)}; \rho_{y_{ij}y_{ik}})\Big\}\\ &\quad - \Big\{\Phi_2(\tau_j^{(c_j-1)}, \tau_k^{(c_k)}; \rho_{y_{ij}y_{ik}}) - \Phi_2(\tau_j^{(c_j-1)}, \tau_k^{(c_k-1)}; \rho_{y_{ij}y_{ik}})\Big\}, \qquad (14)\end{aligned}$$

where $\Phi_2(a, b; \rho)$ is the bivariate cumulative standard normal distribution with correlation ρ evaluated at the point (a, b). The correlation is obtained from the model parameters via

$$\rho_{y_{ij}y_{ik}} = \lambda_j \lambda_k, \qquad (15)$$

for $j = 1, \ldots, (p-1)$ and $k = (j+1), \ldots, p$.

Comparing Equation (8) with Equation (10), we can see that the p-dimensional integral is reduced to all possible pairwise ($j < k$) integrals, which are bivariate normal distribution with closed form solution. This significantly reduces the computational complexity, which is a major advantage of PML.

4. Maximizing the likelihood function

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

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{\theta}), \qquad (16)$$

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

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

$$\sum_{i=1}^{n} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_i) = \boldsymbol{0}, \qquad (17)$$

where

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}; \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n). \qquad (18)$$

and

$$\begin{aligned}\boldsymbol{s}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_i) &= \frac{\partial \ell(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \qquad (19)\\ &= \left(\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y}_i)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y}_i)}{\partial \theta_q}\right). \qquad (20)\end{aligned}$$

For the two-parameter IRT model analyzed in this paper, the log-likelihood function and consequently also the individual score function differs depending on whether we are estimating models via MML or via PML.

4.1. MML likelihood and score functions

Marginal ML explicitly solves the integral from (7) using numerical methods. Assuming the two-parameter IRT model described above, the score vector for each individual can be written as:

$$\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{y}_i) = \left(\frac{\partial \ell(\boldsymbol{\theta}; y_{i1})}{\partial \alpha_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}; y_{ip})}{\partial \alpha_p}, \frac{\partial \ell(\boldsymbol{\theta}; y_{i1})}{\partial \gamma_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}; y_{ip})}{\partial \gamma_p} \right), \tag{21}$$

where the likelihood function for each observation $\ell(\boldsymbol{\theta}; y_{ij})$ can be expressed as

$$\ell(\boldsymbol{\theta}; y_{ij}) = \log \int_{-\infty}^{\infty} f(\boldsymbol{\theta}; y_{ij}|\eta_i) g(\eta_i) d\eta_i \tag{22}$$

Plugging Equation (3) in to Equation (22), we obtain the following functional form:

$$\ell(\boldsymbol{\theta}; y_{ij}) = \int_{-\infty}^{\infty} \left\{ y_{ij}(\alpha_j \eta_i + \gamma_j) - \log(1 + \exp(\alpha_j \eta_i + \gamma_j)) + \log(g(\eta_i)) \right\} d\eta_i. \tag{23}$$

Therefore, components of the score vector from (21) can be written as:

$$\frac{\partial \ell(\boldsymbol{\theta}; y_{ij})}{\partial \alpha_j} = \int_{-\infty}^{\infty} \left\{ y_{ij}\eta_i - \frac{\eta_i \exp(\gamma_j + \alpha_j \eta_i)}{1 + \exp(\gamma_j + \alpha_j \eta_i)} \right\} d\eta_i \tag{24}$$

$$\frac{\partial \ell(\boldsymbol{\theta}; y_{ij})}{\partial \gamma_j} = \int_{-\infty}^{\infty} \left\{ y_{ij} - \frac{\exp(\gamma_j + \alpha_j \eta_i)}{1 + \exp(\gamma_j + \alpha_j \eta_i))} \right\} d\eta_i. \tag{25}$$

These integrals are often approximated via Gauss-Hermite quadrature. If the dimension of ability (η) increases (multidimensional trait), the quadrature procedure becomes infeasible. To avoid this difficulty and to remove the need for quadrature, pairwise maximum likelihood estimation (PML) can be used instead. This is described in the following section.

4.2. PML likelihood and score functions

Maximizing the log-likelihood function in Equation (10) over the parameter $\boldsymbol{\theta}$, we obtain the composite pairwise maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\text{PML}}$. Again, this is equivalent to solving for $\boldsymbol{\theta}$ 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 $\boldsymbol{\tau}$:

$$\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{y}_i) = \left(\frac{\partial \left\{ \sum\limits_{j<k} \ell(\boldsymbol{\theta}; (y_{ij}, y_{ik})) \right\}}{\partial \boldsymbol{\Lambda}}, \frac{\partial \left\{ \sum\limits_{j<k} \ell(\boldsymbol{\theta}; (y_{ij}, y_{ik})) \right\}}{\partial \boldsymbol{\tau}} \right). \tag{26}$$

The elements of the score matrix are *analytical* solutions, requiring no approximation via quadrature. The derivatives are explicitly shown in Appendix A.

Comparing MML with PML, we can see that scores from PML are more easily obtained and less computationally intensive. Thus, we focus on PML in the simulations and analyses below, with similar results holding for MML as demonstrated in Appendix B. In the next section, we describe the scores' use in tests of measurement invariance.

5. Measurement invariance hypothesis test

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 $\boldsymbol{\theta}_i$. The null hypothesis of measurement invariance can then be expressed as all observations arising from a common set of population parameters $\boldsymbol{\theta}_0$

$$H_0 : \boldsymbol{\theta}_i = \boldsymbol{\theta}_0 \quad (i = 1, \ldots, n), \tag{27}$$

versus

$$H_1 : \boldsymbol{\theta}_i = \boldsymbol{\theta}(v_i) \quad (i = 1, \ldots, n), \tag{28}$$

where $\boldsymbol{\theta}(v_i)$ 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:

$$H_1 : \boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}^{(A)} & v_i \leq v \\ \boldsymbol{\theta}^{(B)} & v_i > v. \end{cases} \tag{29}$$

For this hypothesis testing problem, 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

$$LR(v) = -2[\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) - \{\ell(\hat{\boldsymbol{\theta}}^{(A)}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) + \ell(\hat{\boldsymbol{\theta}}^{(B)}; \boldsymbol{y}_{m+1}, \ldots, \boldsymbol{y}_n)\}], \tag{30}$$

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

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:

$$\max_{v \in [\underline{v}, \bar{v}]} LR(v). \tag{31}$$

The asymptotic distribution of this maximum LR statistic is not χ^2; Andrews (1993) showed that, under the null hypothesis in (27), the statistic converges in distribution to some stochastic process. This result is also utilized in the score-based tests discussed below.

6. Score-based tests

In this section, we review the score-based tests' theoretical background and describe a family of test statistics that we can obtain via the theory. Related descriptions can be found in Zeileis and Hornik (2007), Merkle *et al.* (2014), and Wang *et al.* (2014).

6.1. Theoretical background

The score-based tests described here can be viewed as a generalization of the Lagrange multiplier test (e.g., Satorra 1989). The tests utilize score functions such as those derived above for the PML and MML estimation methods, and they are based on theory showing that functions of the scores follow a stochastic process along an auxiliary variable V.

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 a cumulative score as

$$\boldsymbol{B}(t;\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_{(i)}) \qquad (0 \leq t \leq 1), \tag{32}$$

where $\boldsymbol{y}_{(i)}$ represents the observed data vector for ith-largest observation, with ordering determined by the auxiliary variable V. $\hat{\boldsymbol{I}}$ 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, $\boldsymbol{B}(t;\hat{\boldsymbol{\theta}})$ changes at $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$. For $t = 1$ the cumulative score vector always equals $\boldsymbol{0}$, as defined in Equation (17). 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 (32) as $\hat{\boldsymbol{I}}$. 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 $\hat{\boldsymbol{I}} = (1/n)\sum_{i=1}^{n} \boldsymbol{s}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}_{(i)}) \boldsymbol{s}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}_{(i)})^T$.

Hjort and Koning (2002) showed that, under the null hypothesis from (27), $\boldsymbol{B}(t;\hat{\boldsymbol{\theta}})$ converges in distribution to an independent Brownian bridge:

$$\boldsymbol{B}(\cdot;\hat{\boldsymbol{\theta}}) \xrightarrow{d} \boldsymbol{B}^0(\cdot), \tag{33}$$

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

Empirically, the $\boldsymbol{B}(t;\boldsymbol{\theta})$ 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 (17). To obtain scalar test statistics, we summarize the empirical behavior of Equation (32) and compare it to the analogous scalar summary of the Brownian bridge. In the next section, we introduce various summaries of Equation (32) that can serve as test statistics.

6.2. 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 locate 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).

$$DM = \max_{i=1,\dots,n} \max_{j=1,\dots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}|. \tag{34}$$

However, the DM statistic is sub-optimal if many of the parameters change and/or there exist many changing points of V instead of one, because it "wastes" power by only taking the maximum. In such cases, sums across parameters and individuals are more suitable. The Cramèr-von Mises (CvM) statistic falls in this category,

$$CvM = n^{-1} \sum_{i=1,\dots,n} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2. \tag{35}$$

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

$$\max LM = \max_{i=\underline{i},\bar{i}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2. \tag{36}$$

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

$$\begin{aligned} WDM_o &= \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1/2} \max_{j=1,\dots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}|, & (37) \\ \max LM_o &= \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, & (38) \end{aligned}$$

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

$$LM_{uo} = \sum_{l=1,\dots,m} \sum_{j=1,\dots,k} \left(\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_l j} - \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_{l-1} j} \right)^2, \tag{39}$$

where $\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_0 j} = 0$ for all j. This statistic is asymptotically equivalent to the usual, likelihood ratio statistic, and it is advantageous over the LRT from (30) because it requires the estimation of only one model (the null model).

In the following sections, we apply these theoretical results to IRT models. We focus on the two-parameter model where the η_i are assumed to arise from a normal distribution. We focus on the PML estimation for its speed and present comparable MML results in Appendix B.

7. 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.

7.1. 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 (α_3) or intercept parameter (γ_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 (γ_3 or α_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 occured 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 $d = 0, 0.25, 0.5, \ldots, 4$.

For each combination of sample size (n) × violation magnitude (d) × violating parameter × groups (m), 5,000 data sets were generated and tested. In all conditions, we maintained equal sample sizes in each subgroup of the categories m. Statistics from Equations (37) and (38) (both ordinal statistics) were examined, as was the statistic from (39) (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.

7.2. Results

Full simulation results for PML are presented in Figures 1 to 4 (similar results for MML are shown in Appendix B). 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,

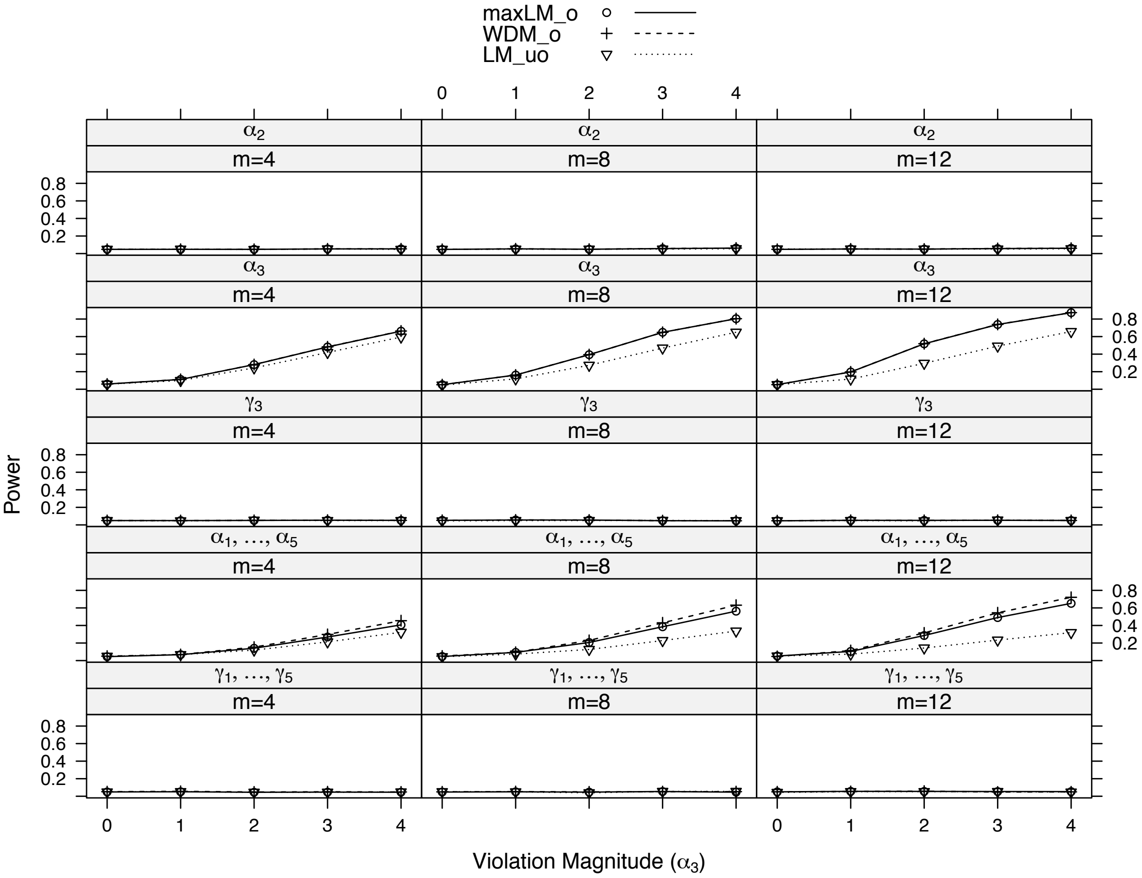

Figure 1: Simulation 1. Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by the PML two-parameter model. The parameter violating measurement invariance is α_3. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

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 α_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 γ_3.

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).

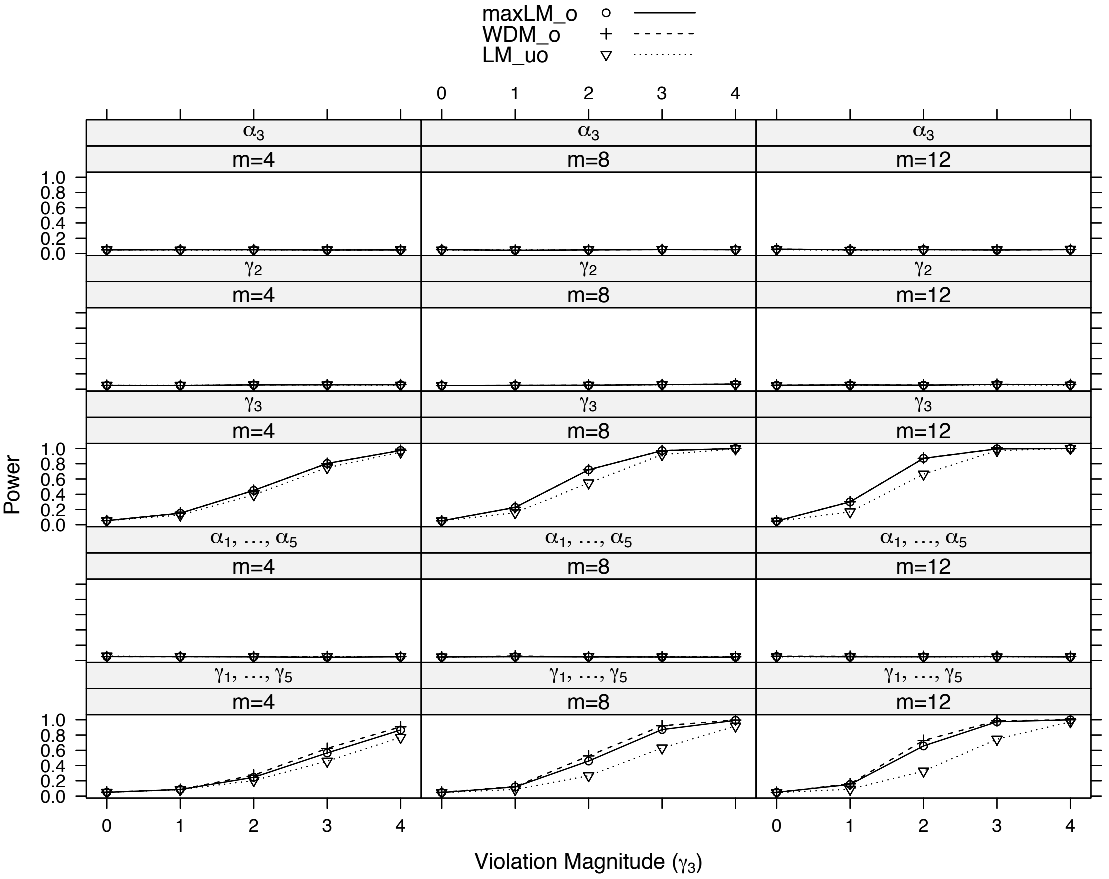

Figure 2: Simulation 1. Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML two-parameter model. The parameter violating measurement invariance is γ_3. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

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 α_3, and Figure 4 displays the results when the violating parameter is γ_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" (we can see it as intercept × 1) than in "interactions" (slope × person parameter η_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 η_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

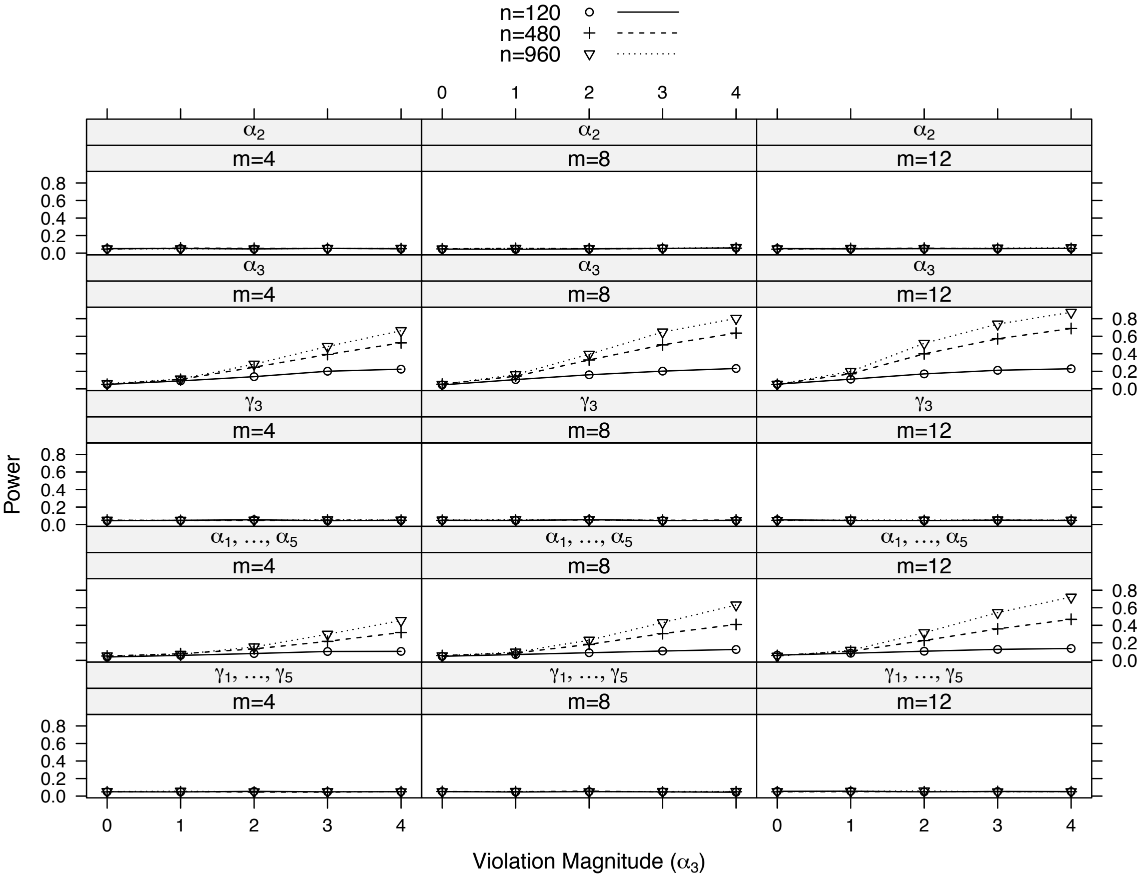

Figure 3: Simulation 1. Simulated power curves for observation 120, 480 and 960 of test statistic WDM_o, across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML two-parameter model. The parameter violating measurement invariance is α_3. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m

slope parameter is attached to the person parameter η_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 (38)) whereas WDM_o only takes the maximum over parameters (see Equation (37)). However, the relative power advantage of using max LM_o 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 viola-

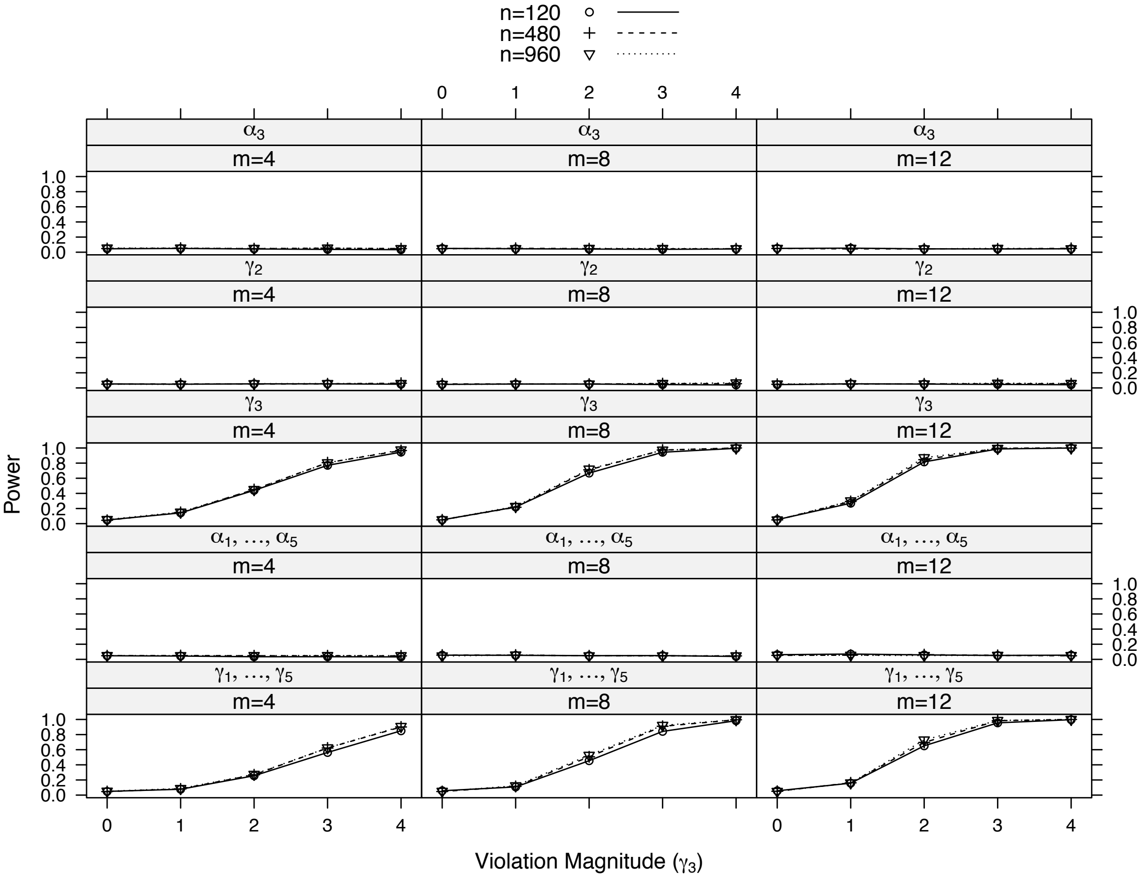

Figure 4: Simulation 1. Simulated power curves for observation 120, 480 and 960 of test statistic WDM_o, across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML two-parameter item response model. The parameter violating measurement invariance is γ_3. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m

tions 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.

8. 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: $\eta_i \sim N(0, 1)$. There is the potential that the hyperdistribution is group specific, however, with $\eta_i^\star \sim N(\mu_{v_i}, \sigma^2_{v_i})$, where v_i is in $1, \ldots, m$. If the hyperparameters change from group to group, then our model can be written as:

$$\begin{aligned} \Phi^{-1}(p_{ij}) &= \gamma_j + \alpha_j \eta_i^\star, && (40) \\ &= \gamma_j + \alpha_j(\mu_{v_i} + \sigma_{v_i}\eta_i), && (41) \\ &= (\gamma_j + \alpha_j \mu_{v_i}) + \sigma_{v_i}\alpha_j\eta_i. && (42) \end{aligned}$$

This shows that, when σ_{v_i} differs across values of v_i, it will look like there are measurement invariance violations in α_j (for all j). Similarly, when μ_{v_i} differs across values of v_i, it will look like there are measurement invariance violations in γ_j (for all j). Further, because μ_{v_i} is no longer 0, changes in α_j will also make it look like there are measurement invariance violations in the γ_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 μ_{v_i} and $\sigma^2_{v_i}$, when there is uncertainty about person abilities. Estimation of these extra parameters will decrease the proposed tests' power, but the extent of decrease is unclear. The extent of power decrease in the proposed test statistics, as compared to traditional statistics, is also unclear. In this section, we conduct two simulations that address these issues.

8.1. 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 $\eta_i \sim N(0, 1)$ 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 $\eta_i \sim N(0, 1)$, while the abilities of students with $V = 5, 6, 7$, or 8 were generated from $\eta_i \sim N(-1, 2)$.

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 μ_{v_i} (with level 1 fixed to zero for identification), $\sigma^2_{v_i}$ (with level 1 fixed to 1 for identification) and the 5 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 μ_{v_i} and 7 variance parameters $\sigma^2_{v_i}$), the sample sizes were increased to $n = 1200, 4800$, and 9600. Measurement invariance violations still occurred in the same places (either α_3 or γ_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) × 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 (37), (38) and (39).

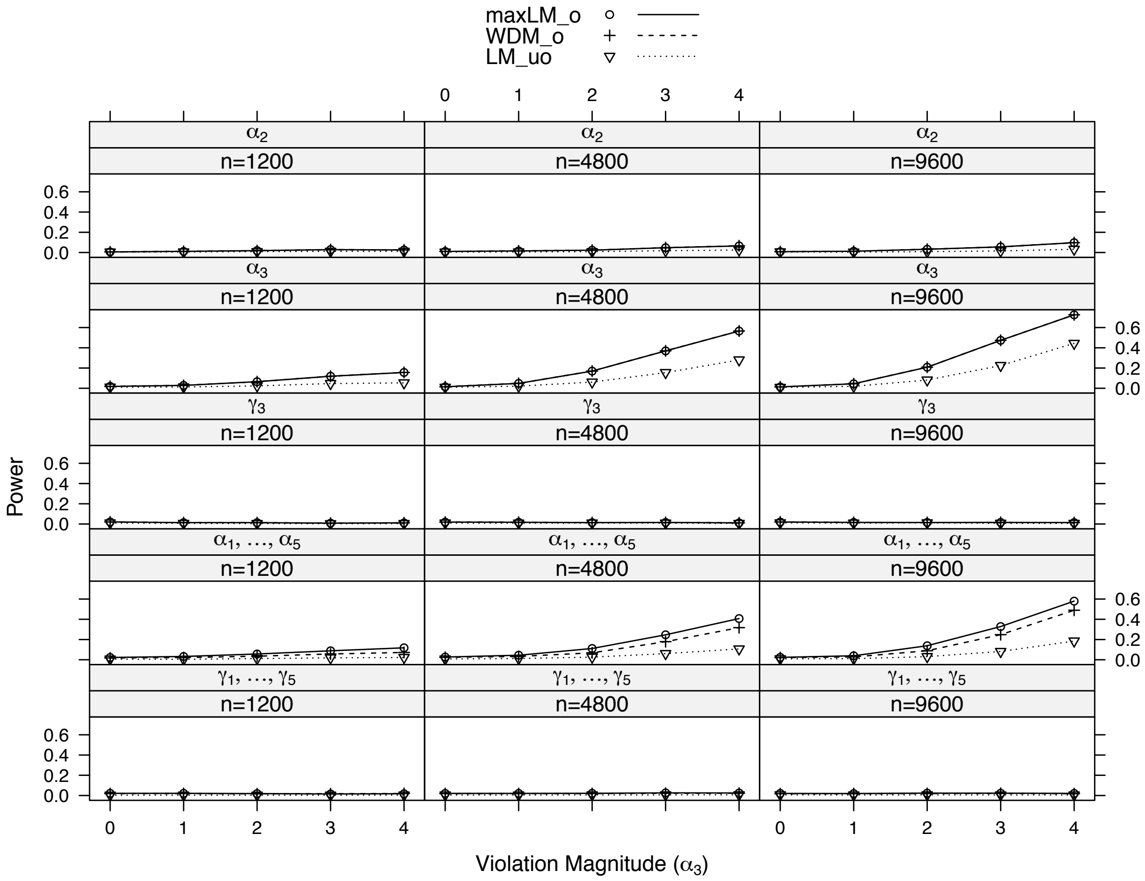

Figure 5: Simulation 2.1. Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting multiple group two-parameter model, without person abilities change in the generation model). The parameter violating measurement invariance is α_3. The number of categories is $m = 8$. Panel labels denote the parameter(s) being tested and sample size.

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

Simulation 2.1

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 α_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 γ_3. In both figures, tests of item 2's parameters are

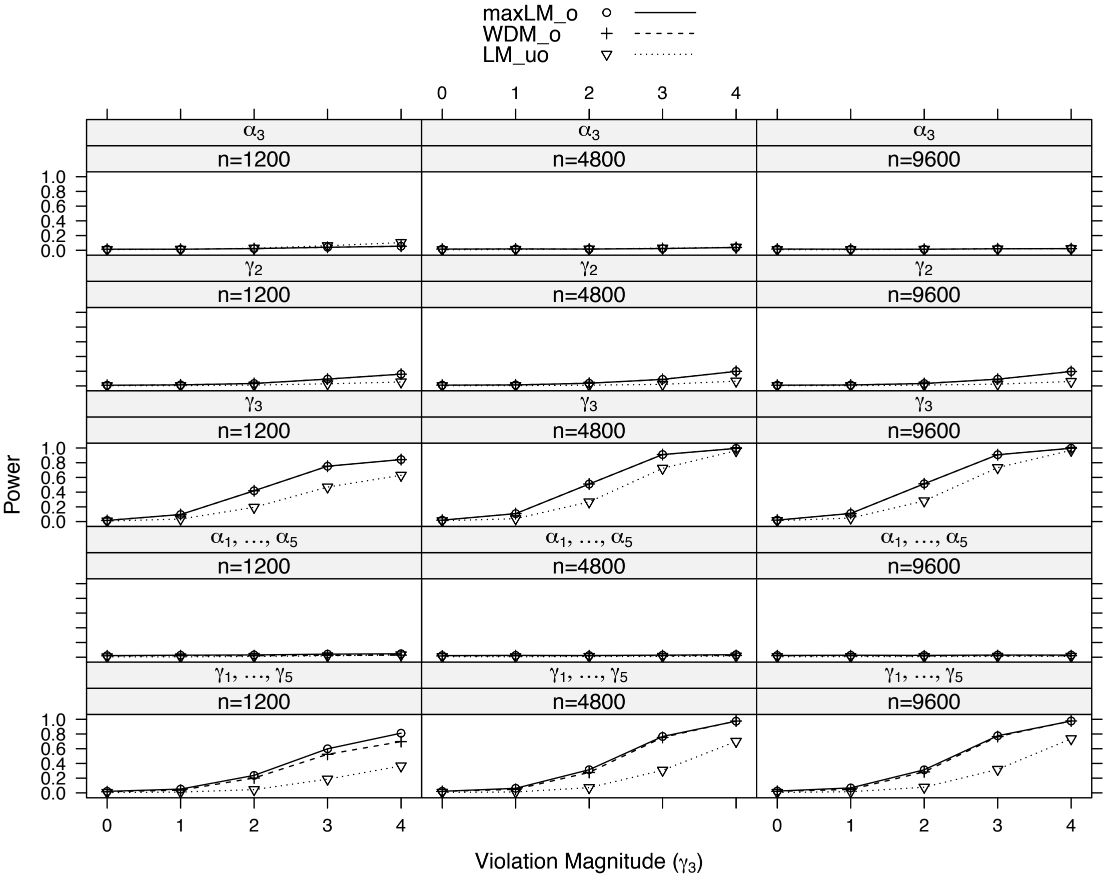

Figure 6: Simulation 2.1. Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting multiple group two-parameter model, without person abilities change in the generation model). The parameter violating measurement invariance is γ_3. The number of categories is $m = 8$. Panel labels denote the parameter(s) being tested and sample size.

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 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.

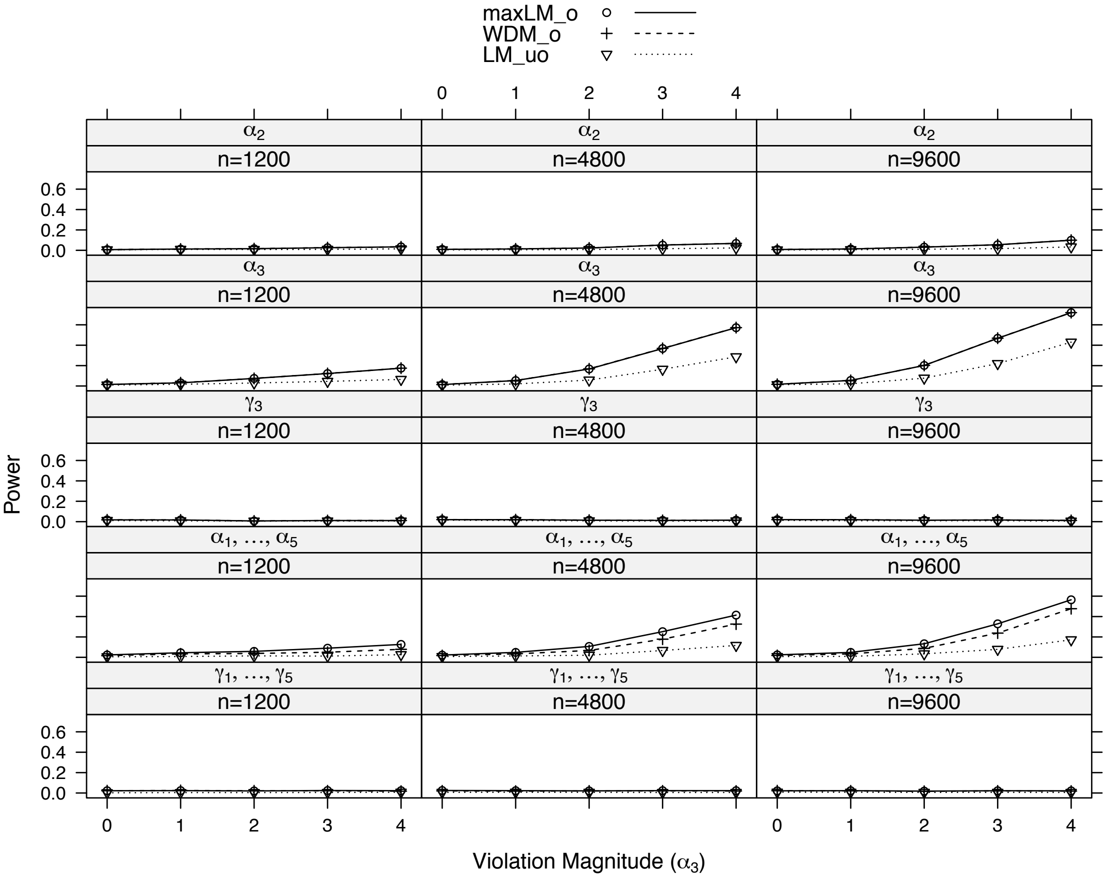

Figure 7: Simulation 2.2. Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting multiple group two-parameter model, with person abilities change in the generation model). The parameter violating measurement invariance is α_3. The number of categories is $m = 8$. Panel labels denote the parameter(s) being tested and sample size.

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

Simulation 2.2 results are presented in Figures 7 and 8, with the same figure and panel arrangements as Simulation 2.1. The results 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

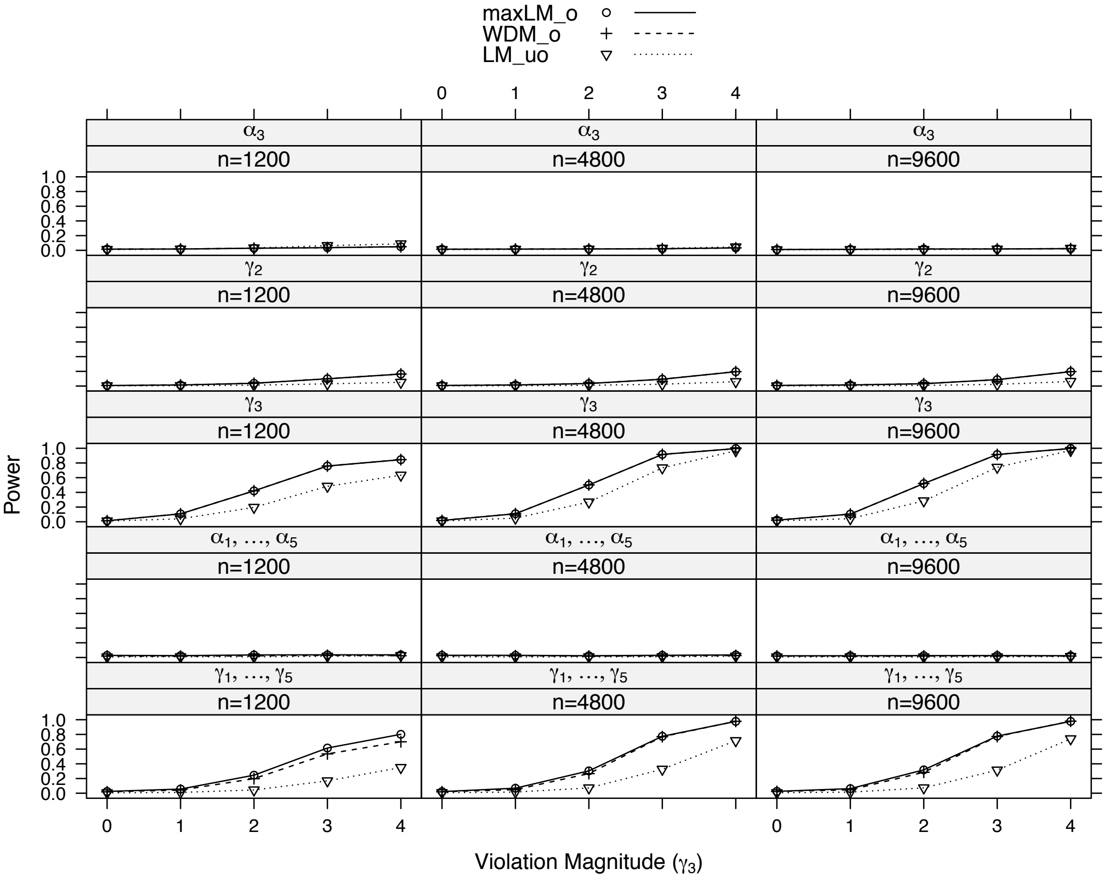

Figure 8: Simulation 2.2. Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting multiple group two-parameter model, with person abilities change in the generation model). The parameter violating measurement invariance is γ_3. The number of categories is $m = 8$. Panel labels denote the parameter(s) being tested and sample size.

multiple group model requires a much larger sample size to obtain reasonable power, this type of model is very necessary in practice when there is uncertainty about changes in person hyperparameters. Otherwise, there will be a serious "false alarm" as illustrated by Equations 40 to 42. It seems that the sample size issue can often be addressed, as IRT researchers often have thousands of students completing their tests. In the following section, we demonstrate the tests' use in a practical situation.

9. Application

We illustrate the tests' application using 20 dichotomously scored mathematics items from the graduation examination developed by the Netherlands National Institute for Educational Measurement (Doolaard 1999; Fox 2010).

9.1. 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 20 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.

From the simulations, we saw that accounting for changes in person hyperparameters is crucial to avoid "false alarms;" however, increasing the number of person hyperparameters will decrease power. Therefore, we start with a 2-parameter item response model where the person hyperparameters μ_1 and σ_1^2 (for level 1) are fixed to 0 and 1, while the hyperparameters in other levels are freely estimated but constrained to be equal. We can then test whether the hyperparameters are equal across levels and, if not, we can use the test results to freely estimate hyperparameters across levels.

9.2. Results

We describe the results in two sections, one for the initial examination of fluctuations in the hyperparameters, and one for a second model that frees certain hyperparameters.

Testing the hyperparameters

Results representing the statistics' fluctuations across SES level are shown in Figure 9. The first column displays the fluctuation process associated with LM_o for testing the 18 items' slopes (first row), the 18 items' intercept (second row), the person mean parameter (third row), and the person variance parameter (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 value of Equations (38) and (37) for each SES level, with the dashed horizontal line being the 5% critical value. If the solid line crosses the critical value, then it is evidence that the corresponding parameter fluctuates across levels of SES. Because the final level's statistics always equal zero (see Equation (17)), 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 will 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 process, the second model should estimate a separate person mean parameter for each SES level and should estimate a separate person variance parameter for the middle levels (level 2 – level 4) and for the extreme levels (at and after level 5). The result for testing parameters in this freed-hyperparameters model is described in the next section.

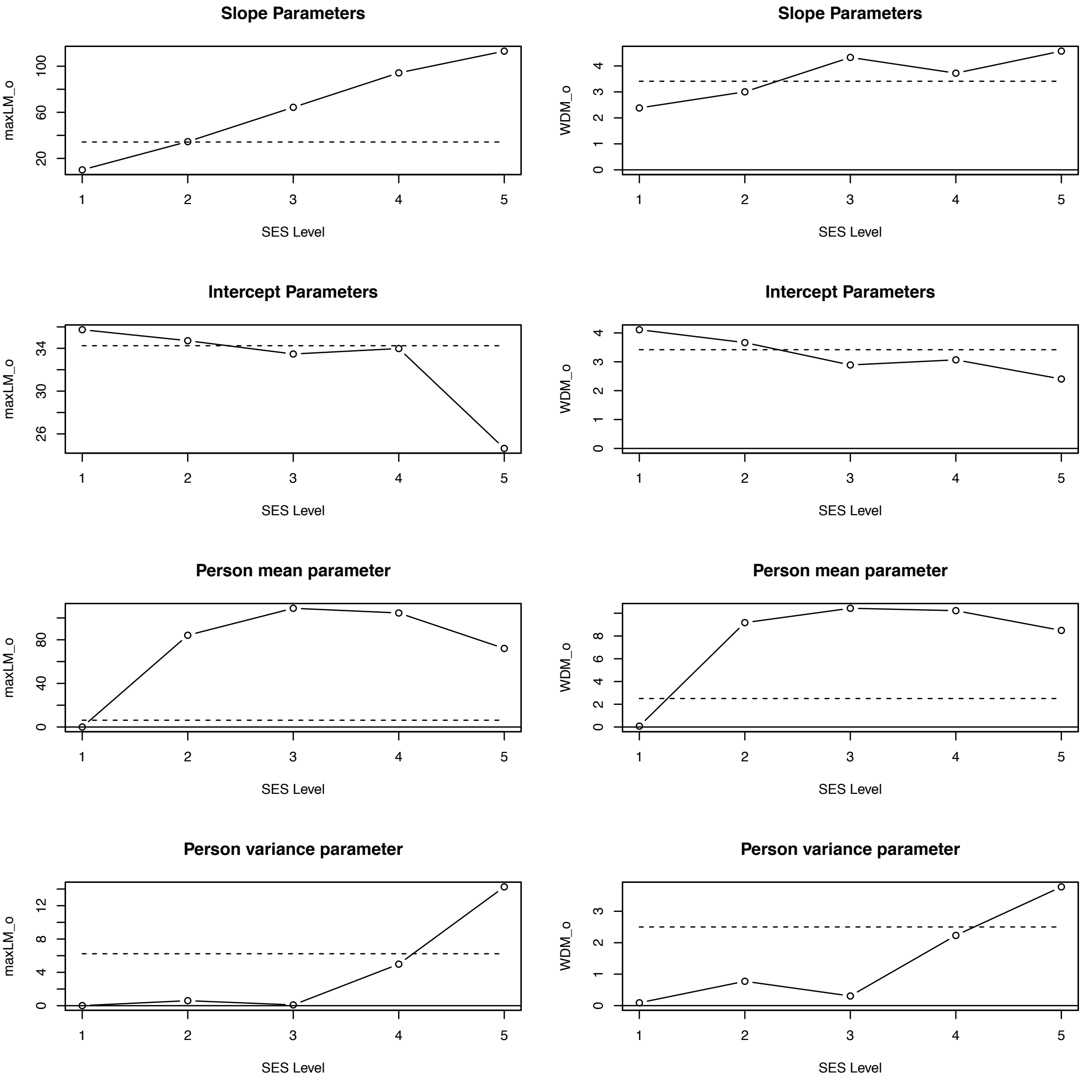

Figure 9: Empirical fluctuation processes of the max LM_o statistic (first column) and WDM_o (second column) for slope parameters (first row), intercept parameters (second row), person mean parameter (third row) and person variance parameter (fourth row), using testing-hyperparameters model

Freed hyperparameters

In estimating a separate μ_{v_i} for each of the six SES groups (with first level being fixed to 0 for identification) and two separate σ^2s for the middle level and extreme levels, we obtain the results shown in Figure 10. The panel arrangements are the same as Figure 9.

Figure 10 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. Further, the estimated μ_{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, σ^2 for the middle SES levels (level 2–level 4) and extreme SES levels (level 5–level 6) are 1.14 and 1.37, with the lowest SES

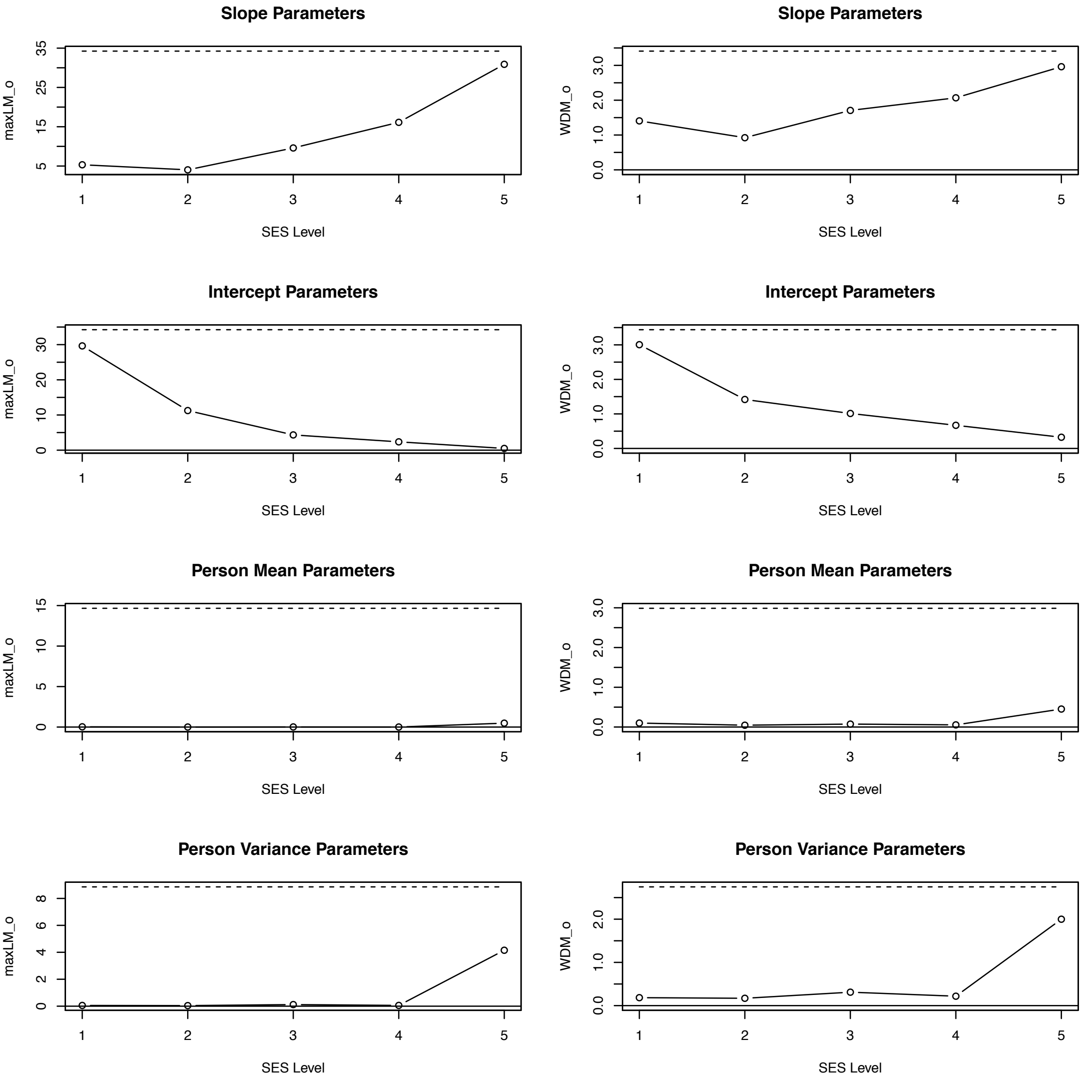

Figure 10: Empirical fluctuation processes of the max LM_o statistic (first column) and WDM_o (second column) for slope parameters (first row), intercept parameters (second row), person mean parameters (third row) and person variance parameters (fourth row), using freed-hyperparameter model

level having a fixed variance of 1.

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.

10. 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 enhances 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.

Multiple category responses

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).

Multidimensional IRT

Multidimensionality is one possible cause of DIF (Millsap 2012). However, it is difficult to test this hypothesis in IRT models. The challenge is caused by the integration described in the Estimation section. 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 invariance in larger datasets.

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 modeling (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.

10.1. Further development

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 intercept parameters, the current tests can still be applied. However, when there are multiple violating slope parameters, we need to use the tests in a recursive way. This would 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) (also known as "purification"), which is implemented in R packages **mirt** (Chalmers 2012) and **difR** (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.

Nonlinear parameter constraints

In this paper, we constrained the first group mean and variance to 0 and 1 to identify the model. This is appropriate when the auxiliary variable is ordinal because it allows us to observe group ability/variance change along the ordinal auxiliary variable. However, it may be better to use "sum" constraints when the auxiliary variable is categorical because we do not need any pre-defined order for group ability or anchor items. Verhagen, Levy, Millsap, and Fox (2015) constrained the sum of all intercept parameters to be zero (in a one parameter IRT model) to avoid the need for defining anchor items or assuming group ability (i.e. fixing one group ability parameter). We can extend these constraints to the slopes of a two-parameter model, requiring that the squared slope parameters sum to 1. We are currently studying use of these nonlinear parameter constraints in tandem with score tests.

Application in multilevel models

In educational research settings, students' responses to items are often nested in classes, schools, or states. Therefore, multilevel models are generally applied in this area. Score-based tests only rely on the derivative of each individual's likelihood function so that, as long as the individual derivative (analytical or approximation) can be specified, score-based 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.

10.2. 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 2013), version 3.3.0, employing the add-on package **lavaan** 0.5-17 (Rosseel 2012) for fitting of the factor analysis models and **strucchange** 1.5-2 (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 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`.

Ayala RJD (2009). *The Theory and Practice of Item Response Theory*. Guilford Press.

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`.

den Noortgate WV, Boeck PD (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`.

Doolaard S (1999). "Schools in Change or Schools in Chains." *Unpublished doctoral dissertation, University of Twente, The Netherlands.*

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 *Rasch Models*, pp. 15–38. Springer-Verlag.

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.

Fox JP (2010). *Bayesian Item Response Modeling: Theory and Applications*. Springer-Verlag. `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.

Hjort NL, Koning A (2002). "Tests for Constancy of Model Parameters over Time." *Nonparametric Statistics*, **14**, 113–132. doi:10.1080/10485250211394.

Holland PW, Thayer DT (1988). "Differential Item Performance and the Mantel-Haenszel Procedure." *Test Validity*, pp. 129–145.

Jöreskog KG (1990). "New Developments in LISREL: Analysis of Ordinal Variables Using Polychoric Correlations and Weighted Least Squares." *Quality & Quantity*, **24**(4), 387–404. doi:10.1007/bf00152012.

Jöreskog KG, Moustaki I (2001). "Factor Analysis of Ordinal Variables: A Comparison of Three Approaches." *Multivariate Behavioral Research*, **36**(3), 347–387. doi:10.1207/s15327906347-387.

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.

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.

Liu J (2007). *Multivariate Ordinal Data Analysis with Pairwise Likelihood and Its Extension to SEM*. Ph.D. thesis, University of California, Los Angeles.

Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems*. Routledge. doi:10.4324/9780203056615.

Magis D, Beland S, Raiche G (2015). *difR: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF)*. R package version 4.6.

Magis D, Béland S, Tuerlinckx F, Boeck PD (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 (2012). "Angoff's Delta Method Revisited: Improving DIF Detection under Small Samples." *British Journal of Mathematical and Statistical Psychology*, **65**(2), 302–321. doi:10.1111/j.2044-8317.2011.02025.x.

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, NJ.

Millsap RE (2012). *Statistical Approaches to Measurement Invariance*. Routledge. doi:10.4324/9780203821961.

Muraki E (1992). "A Generalized Partial Credit Model: Application of an EM Algorithm." *Applied Psychological Measurement*, **16**, 159–176. doi:10.1177/014662169201600206.

Muthén B (1984). "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators." *Psychometrika*, **49**(1), 115–132. doi:10.1007/bf02294210.

Olsson U (1979). "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient." *Psychometrika*, **44**(4), 443–460. doi:10.1007/bf02296207.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rijmen F, Tuerlinckx F, Boeck PD, Kuppens P (2003). "A Nonlinear Mixed Model Framework for Item Response Theory." *Psychological Methods*, **8**(2), 185. doi:10.1037/1082-989x.8.2.185.

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/bf02290599.

Satorra A (1989). "Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach." *Psychometrika*, **54**, 131–151. doi:10.1007/bf02294453.

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. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Verhagen J, Levy R, Millsap RE, Fox JP (2015). "Evaluating Evidence for Invariant Items: A Bayes Factor Applied to Testing Measurement Invariance in IRT Models." *Journal of Mathematical Psychology*. 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.

A. Model scores

This appendix contains information about how to derive the score vectors' components in Equation (26). It is organized by deriving the score sub-vector w.r.t. $\boldsymbol{\tau}$ and w.r.t. $\boldsymbol{\Lambda}$. Details are adapted from Katsikatsou *et al.* (2012).

A.1. Score sub-vector w.r.t. $\boldsymbol{\tau}$

The elements of the sub-vector $\frac{\partial\left\{\sum_{j<k} \ell(\boldsymbol{\theta};(y_{ij},y_{ik}))\right\}}{\partial\boldsymbol{\tau}}$ are the first derivatives with respect to thresholds $\frac{\partial\left\{\sum_{j<k} \ell(\boldsymbol{\theta};(y_{ij},y_{ik}))\right\}}{\partial\tau_j^{(c_j)}}$ and are given as below:

$$\frac{\partial\left\{\sum\limits_{j<k} \ell(\boldsymbol{\theta};(y_{ij},y_{ik}))\right\}}{\partial\tau_j^{(c_j)}} = \sum_{c_j=1}^{2}\left(\frac{1}{\pi_{y_{ij}y_{ik}}^{(c_jc_k)}} - \frac{1}{\pi_{y_{ij}y_{ik}}^{((c_j-1)c_k)}}\right)\frac{\partial\pi_{y_{ij}y_{ik}}^{(c_jc_k)}}{\partial\tau_j^{c_j}}, \tag{43}$$

where

$$\frac{\partial\pi_{y_{ij}y_{ik}}^{(c_jc_k)}}{\partial\tau_j^{(c_j)}} = \phi_1(\tau_j^{(c_j)})\left[\Phi_1\left(\frac{\tau_k^{(c_k)} - \rho_{y_{ij}y_{ik}}\tau_j^{(c_j)}}{\sqrt{1-\rho^2_{y_{ij}y_{ik}}}}\right) - \Phi_1\left(\frac{\tau_k^{(c_k-1)} - \rho_{y_{ij}y_{ik}}\tau_j^{(c_j)}}{\sqrt{1-\rho^2_{y_{ij}y_{ik}}}}\right)\right], \tag{44}$$

where ϕ_1 and Φ_1 are the standard univariate normal density and cumulative distribution, respectively.

A.2. Score sub-vector w.r.t. $\boldsymbol{\Lambda}$

The scores with respect to $\boldsymbol{\Lambda}$ are obtained via chain rules.

$$\begin{align} \frac{\partial\left\{\sum\limits_{j<k} \ell(\boldsymbol{\theta};(y_{ij},y_{ik}))\right\}}{\partial\boldsymbol{\Lambda}} &= \frac{\partial\left\{\sum\limits_{j<k} \ell(\boldsymbol{\theta};(y_{ij},y_{ik}))\right\}}{\partial\rho_{y_{ij}y_{ik}}}\frac{\partial\rho_{y_{ij}y_{ik}}}{\partial\boldsymbol{\Lambda}} \tag{45}\\ \tag{46} \end{align}$$

The partial derivative with respect to $\rho_{y_{ij}y_{ik}}$ is

$$\frac{\partial\left\{\sum\limits_{j<k} \ell(\boldsymbol{\theta};(y_{ij},y_{ik}))\right\}}{\partial\rho_{y_jy_k}} = \sum_{c_j=1}^{2}\sum_{c_k=1}^{2}\frac{1}{\pi_{y_{ij}y_{ik}}^{(c_jc_k)}}\frac{\partial\pi_{y_{ij}y_{ik}}^{(c_{ij}c_{ik})}}{\partial\rho_{y_{ij}y_{ik}}}, \tag{47}$$

where $\frac{\partial\pi_{y_{ij}y_{ik}}^{(c_jc_k)}}{\partial\rho_{y_{ij}y_{ik}}}$ is given in Olsson (1979) as

$$\begin{align} \frac{\partial\pi_{y_{ij}y_{ik}}^{(c_jc_k)}}{\partial\rho_{y_{ij}y_{ik}}} = &\left\{\phi_2(\tau_j^{(c_j)},\tau_k^{(c_k)};\rho_{y_{ij}y_{ik}}) - \phi_2(\tau_j^{(c_j)},\tau_k^{(c_k-1)};\rho_{y_{ij}y_{ik}})\right\} \\ &- \left\{\phi_2(\tau_j^{(c_j-1)},\tau_k^{(c_k)};\rho_{y_{ij}y_{ik}}) - \phi_2(\tau_j^{(c_j-1)},\tau_k^{(c_k-1)};\rho_{y_{ij}y_{ik}})\right\}, \tag{48} \end{align}$$

where ϕ_2 is the standard bivariate normal density.

The partial derivative of $\rho_{y_{ij}y_{ik}}$ with respect to $\mathbf{\Lambda}$ can be obtained from a chain rule. The final form of the derivatives is:

$$\frac{\partial \rho_{y_{ij}y_{ik}}}{\partial \mathbf{\Lambda}} = \lambda_k \frac{\partial \lambda_j}{\partial \mathbf{\Lambda}} + \lambda_j \frac{\partial \lambda_k}{\partial \mathbf{\Lambda}}, \tag{49}$$

where $\frac{\partial \Lambda_j}{\partial \mathbf{\Lambda}}$, $\frac{\partial \lambda_k}{\partial \mathbf{\Lambda}}$ are matrices of zeros and ones. The dimensions are determined by $\mathbf{\Lambda}$. The person hyperparameters' score derivation is presented in Liu (2007), Appendix 9.3.

B. MML results from Simulation 1

This appendix demonstrates Simulation 1 results when we fit models via MML, instead of PML. The figures are arranged in the same way as those in the Simulation 1 results section. Figures 11 and 12 display power differences among statistics. Figures 13 and 14 display power differences among sample sizes. Results are similar to those observed for PML.

Affiliation:

Ting Wang, Edgar C. Merkle
Department of Psychological Sciences
University of Missouri
E-mail: twb8d@mail.missouri.edu, merklee@missouri.edu

Carolin Strobl
Department of Psychology
University of Zurich
E-mail: c.strobl@psychologie.uzh.ch

Achim Zeileis
Department of Statistics
Universität Innsbruck
E-mail: Achim.Zeileis@R-project.org

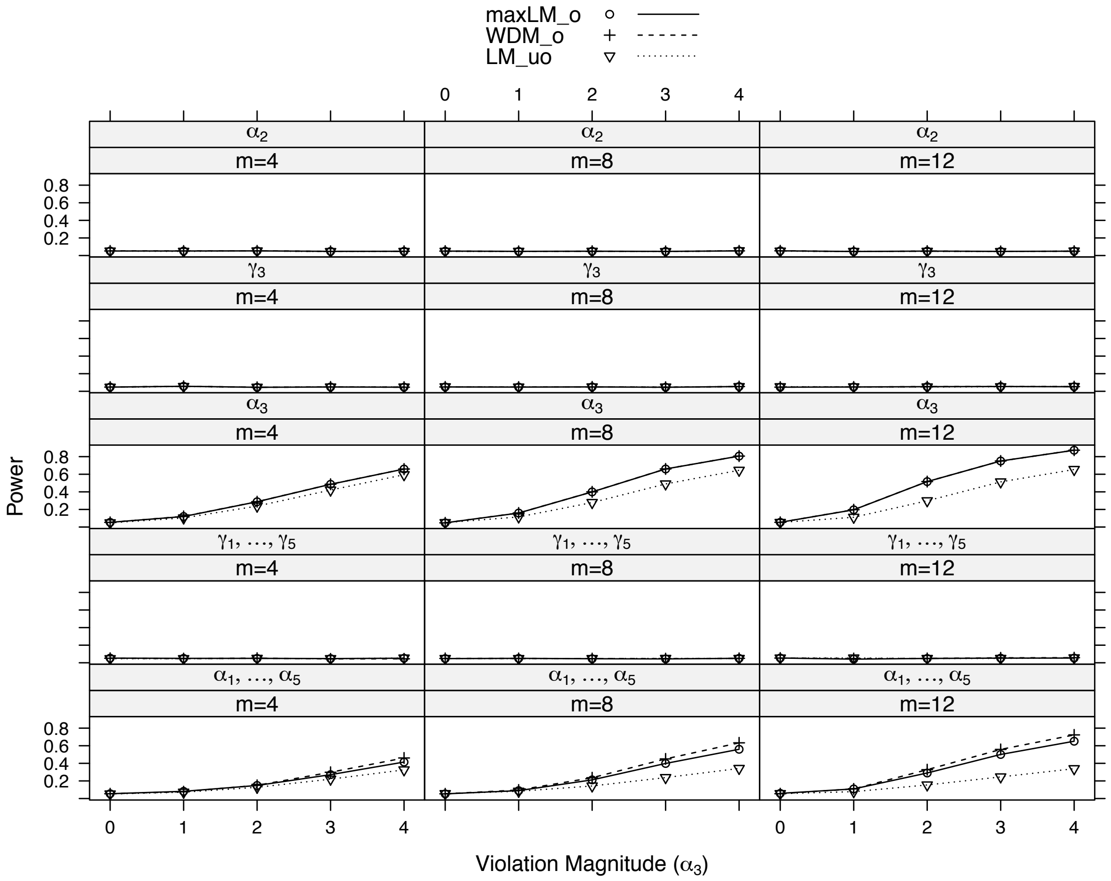

Figure 11: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is α_3. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

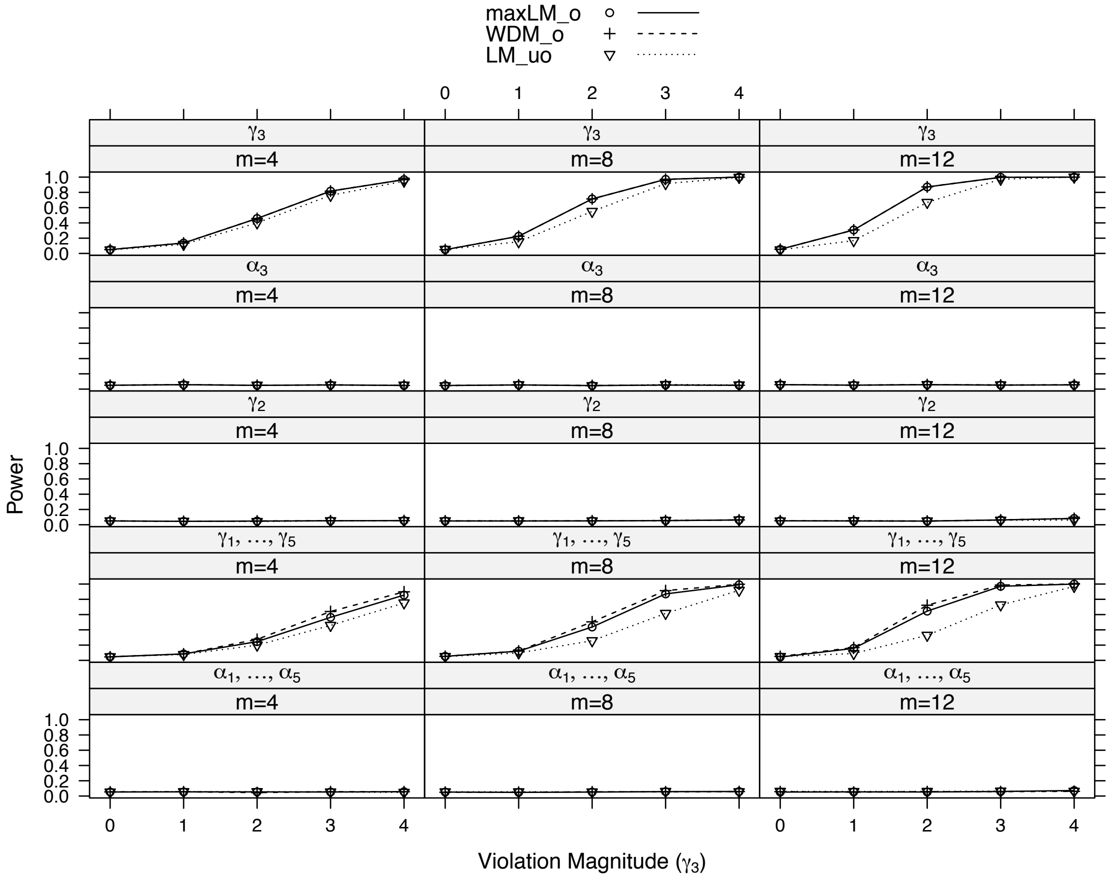

Figure 12: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is γ_3. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

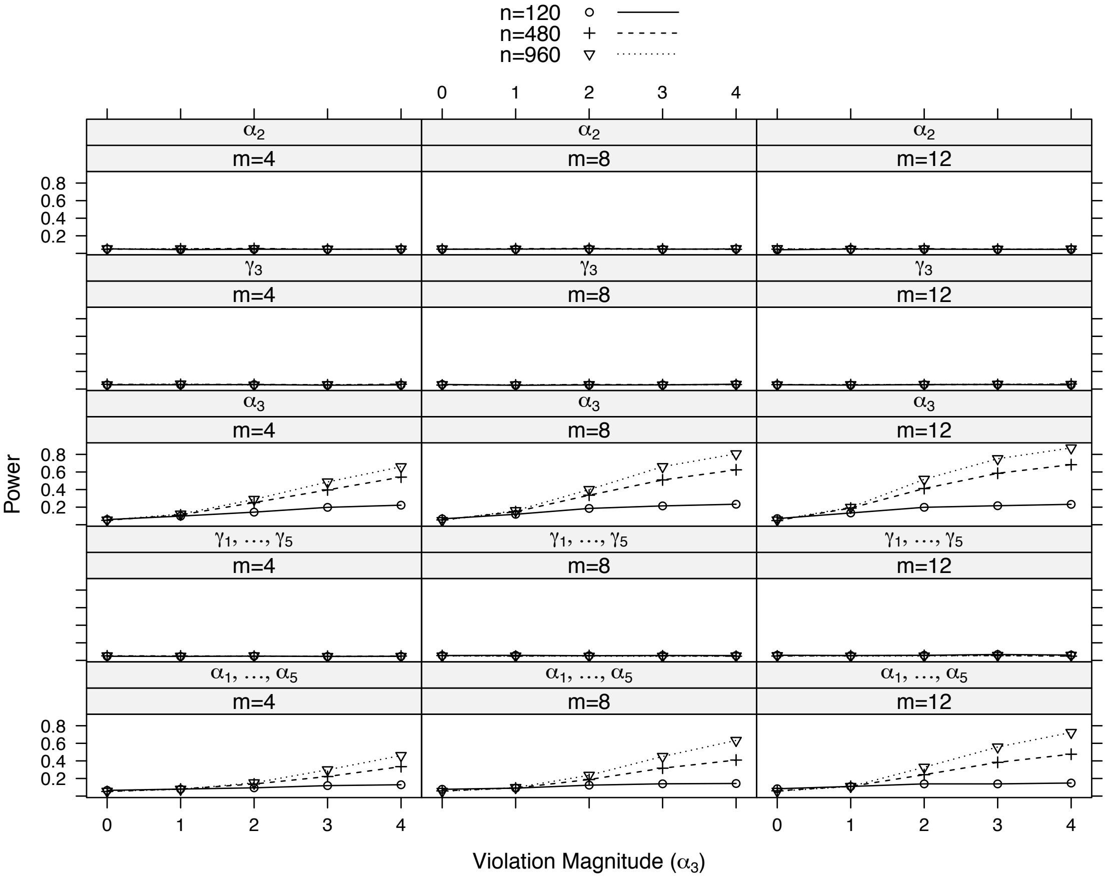

Figure 13: Simulated power curves for observation 120, 480 and 960 of test statistic WDM_o, across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is α_3. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

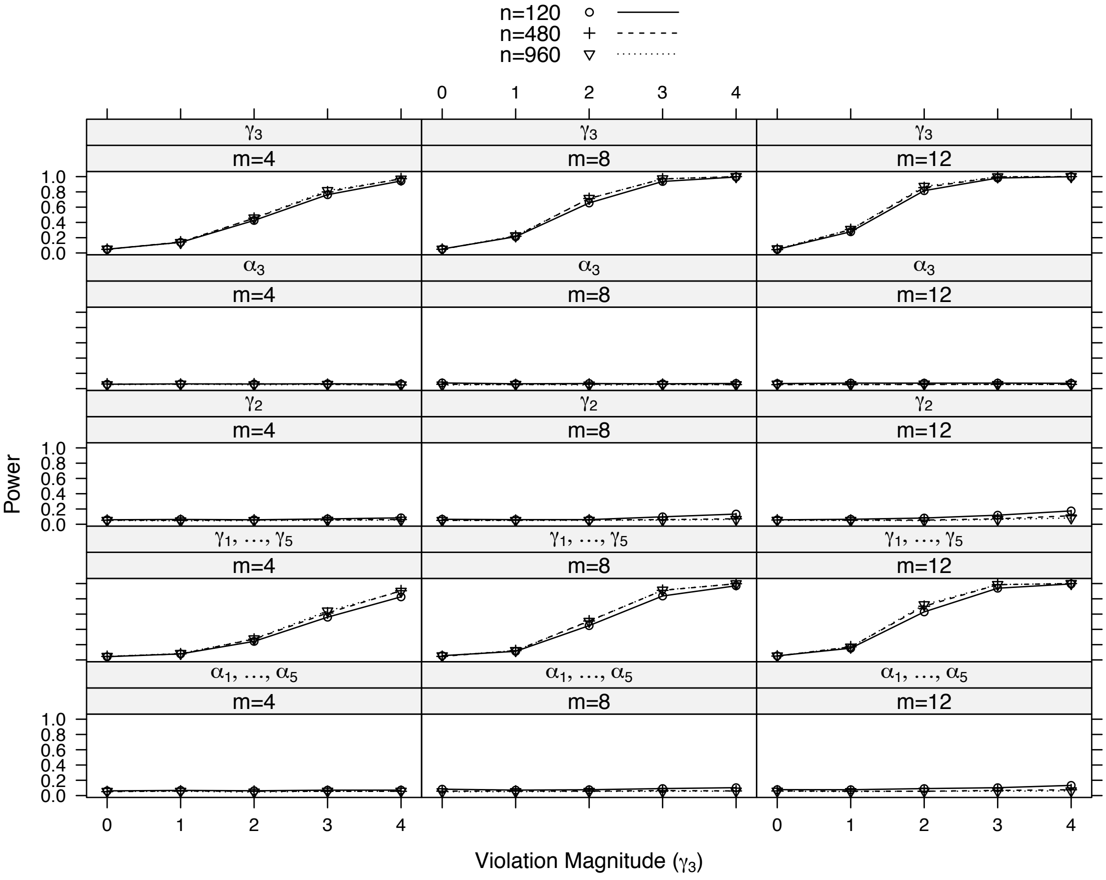

Figure 14: Simulated power curves for observation 120, 480 and 960 of test statistic WDM_o, across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is γ_3. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2016-05 **Ting Wang, Carolin Strobl, Achim Zeileis, Edgar C. Merkle:** Score-based tests of differential item functioning in the two-parameter model

2016-04 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Non-homogeneous boosting for predictor selection in ensemble post-processing

2016-03 **Dietmar Fehr, Matthias Sutter:** Gossip and the efficiency of interactions

2016-02 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Rankings and risk-taking in the finance industry

2016-01 **Sibylle Puntscher, Janette Walde, Gottfried Tappeiner:** Do methodical traps lead to wrong development strategies for welfare? A multilevel approach considering heterogeneity across industrialized and developing countries

2015-16 **Niall Flynn, Christopher Kah, Rudolf Kerschbamer:** Vickrey Auction vs BDM: Difference in bidding behaviour and the impact of other-regarding motives

2015-15 **Christopher Kah, Markus Walzl:** Stochastic stability in a learning dynamic with best response to noisy play

2015-14 **Matthias Siller, Christoph Hauser, Janette Walde, Gottfried Tappeiner:** Measuring regional innovation in one dimension: More lost than gained?

2015-13 **Christoph Hauser, Gottfried Tappeiner, Janette Walde:** The roots of regional trust

2015-12 **Christoph Hauser:** Effects of employee social capital on wage satisfaction, job satisfaction and organizational commitment

2015-11 **Thomas Stöckl:** Dishonest or professional behavior? Can we tell? A comment on: Cohn et al. 2014, Nature 516, 86-89, "Business culture and dishonesty in the banking industry"

2015-10 **Marjolein Fokkema, Niels Smits, Achim Zeileis, Torsten Hothorn, Henk Kelderman:** Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees

2015-09 **Martin Halla, Gerald Pruckner, Thomas Schober:** The cost-effectiveness of developmental screenings: Evidence from a nationwide programme

2015-08 **Lorenz B. Fischer, Michael Pfaffermayr:** The more the merrier? Migration and convergence among European regions

2015-07 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Cooperation and discrimination within and across language borders: Evidence from children in a bilingual city

2015-07 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Cooperation and discrimination within and across language borders: Evidence from children in a bilingual city

2015-06 **Martin Geiger, Wolfgang Luhan, Johann Scharler:** When do Fiscal Consolidations Lead to Consumption Booms? Lessons from a Laboratory Experiment

2015-05 **Alice Sanwald, Engelbert Theurl:** Out-of-pocket payments in the Austrian healthcare system - a distributional analysis

2015-04 **Rudolf Kerschbamer, Matthias Sutter, Uwe Dulleck:** How social preferences shape incentives in (experimental) markets for credence goods *forthcoming in Economic Journal*

2015-03 **Kenneth Harttgen, Stefan Lang, Judith Santer:** Multilevel modelling of child mortality in Africa

2015-02 **Helene Roth, Stefan Lang, Helga Wagner:** Random intercept selection in structured additive regression models

2015-01 **Alice Sanwald, Engelbert Theurl:** Out-of-pocket expenditures for pharmaceuticals: Lessons from the Austrian household budget survey

University of Innsbruck

Working Papers in Economics and Statistics

2016-05

Ting Wang, Carolin Strobl, Achim Zeileis, Edgar C. Merkle

Score-based tests of differential item functioning in the two-parameter model

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 people. 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 estimated via 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.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)