Testing for measurement invariance with respect to an ordinal variable

Edgar C. Merkle
University of Missouri

Jinyan Fan
Auburn University

Achim Zeileis
Universität Innsbruck

Author Note

This work was supported by National Science Foundation grant SES-1061334. Portions of this work were presented at 2012 meeting of the Psychometric Society. The authors thank Yves Rosseel for lavaan asssistance, along with the associate editor and three anonymous reviewers for comments that improved the paper. Correspondence to Edgar C. Merkle, Department of Psychological Sciences, University of Missouri, Columbia, MO 65211. Email: `merklee@missouri.edu`.

Abstract

Researchers are often interested in testing for measurement invariance with respect to an ordinal auxiliary variable such as age group, income class, or school grade. In a factor-analytic context, these tests are traditionally carried out via a likelihood ratio test statistic comparing a model where parameters differ across groups to a model where parameters are equal across groups. This test neglects the fact that the auxiliary variable is ordinal, and it is also known to be overly sensitive at large sample sizes. In this paper, we propose test statistics that explicitly account for the ordinality of the auxiliary variable, resulting in higher power against "monotonic" violations of measurement invariance and lower power against "non-monotonic" ones. The statistics are derived from a family of tests based on stochastic processes that have recently received attention in the psychometric literature. The statistics are illustrated via an application involving real data, and their performance is studied via simulation.

Testing for measurement invariance with respect to an ordinal variable

The study of measurement invariance and differential item functioning (DIF) has received considerable attention in the psychometric literature (see, e.g., Millsap, 2011 for a thorough review). A set of psychometric scales X is defined to be measurement invariant with respect to an auxiliary variable V if (Mellenbergh, 1989)

$$f(x_i|t_i, v_i, \dots) = f(x_i|t_i, \dots), \tag{1}$$

where T is the latent variable that the scales measure, f is the model's distributional form, the i subscript refers to individual cases, capital letters signify random variables, and lowercase letters signify realizations of the variables. If the above equation does not hold, then a measurement invariance violation is said to exist. We focus here on situations where $f()$ is the probability density function of X, and the measurement invariance violation occurs because the model parameters are unequal across individuals (and related to V).

As a concrete example of the study of measurement invariance, consider a situation where X includes "high stakes" tests of ability and V is ethnicity. One's ethnicity should be unrelated to the measurement parameters within $f()$, and this expectation can be studied by fitting the model and examining whether or not measurement parameters vary across different ethnicities. Statistical tools that can be used to carry out this study include likelihood ratio tests, Lagrange multiplier tests, and Wald tests (e.g., Satorra, 1989). These tools have greatly aided in the development of improved, "fairer" psychometric tests and scales.

Along with categorical variables such as ethnicity, researchers are often interested in studying measurement invariance with respect to ordinal V. Such variables can arise from multiple choice surveys, where continuous variables such as age or income are binned into a small number of categories. Alternatively, the variables may arise from gross, qualitative assessments of a particular measure of interest, where individuals may be categorized as having a "low," "medium," or "high" level of the variable of interest. While these variables are relatively easy to find in the literature, there exist very few psychometric methods that specifically account for the fact that V is ordinal. More often, V is treated as categorical so that the traditional tests can be applied. Additionally, if there are many levels, then V may also be treated as continuous. The goals of this paper are to propose two test statistics that explicitly treat V as ordinal and to show that the statistics possess good properties for use in practice.

The test statistics proposed here are derived from a family of tests that were recently applied to the study of measurement invariance in psychometric models (Merkle & Zeileis, 2013; Strobl, Kopf, & Zeileis, 2013). In the following section, we provide an overview of the family and describe the proposed statistics in detail. Subsequently, we report on the results of two simulation studies designed to compare the proposed test statistics to existing tests of measurement invariance. Moreover, we illustrate the proposed statistics using psychometric data on scales purported to measure youth gratitude. Finally, we provide some detail on the tests' use in practice.

Measurement Invariance

In studying measurement invariance, we consider situations where a p-dimensional variable X with observations $\boldsymbol{x}_i, i = 1, \ldots, n$ is described by a model with density $f(\boldsymbol{x}_i; \boldsymbol{\theta})$ and associated joint log-likelihood

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \;=\; \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i) \;=\; \sum_{i=1}^{n} \log f(\boldsymbol{x}_i; \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\theta}$ is some k-dimensional parameter vector that characterizes the distribution.

Tests of measurement invariance are essentially tests of the assumption that all individuals arise from the same parameter vector $\boldsymbol{\theta}$. Thus, a hypothesis of measurement invariance can be written as

$$H_0 : \; \boldsymbol{\theta}_i = \boldsymbol{\theta}_0, \quad (i = 1, \ldots, n), \tag{3}$$

where $\boldsymbol{\theta}_i$ reflects the parameter vector for individual i (and modifications for subsets of $\boldsymbol{\theta}$ are immediate). The most general alternative hypothesis related to V may then be written as

$$H_1^* : \; \boldsymbol{\theta}_i = \boldsymbol{\theta}_{v_i}, \tag{4}$$

stating that the parameter vector differs for every unique realization of V. This alternative is commonly employed when V is categorical. In these situations, the likelihood ratio test (LRT) compares a model where parameters are restricted across groups (i.e., across values of V) to a model where parameters are free across groups; the exact parameter values within each group are completely unrestricted. However, in situations where V is ordinal or continuous, (4) includes non-monotonic violations of measurement invariance. This allows for instances where, e.g., the parameter values initially increase with V and then decrease, or where just one or two "middle" levels of V differ from the rest. Researchers typically do not expect such a result when testing measurement invariance w.r.t. ordinal or continuous V, and researchers often cannot interpret such violations. Monotonic parameter changes w.r.t. V are of much more interest in these situations, with the simplest type of change given by the alternative hypothesis

$$H_2^* : \; \boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}^{(A)} & \text{if } v_i \leq \nu, \\ \boldsymbol{\theta}^{(B)} & \text{if } v_i > \nu, \end{cases} \tag{5}$$

where ν is a threshold dividing individuals into two groups based on V. This alternative is implicitly employed in "median split" analyses, where ν is given as the sample median of V. The threshold ν is usually unknown, however, so it is generally of interest to test (5) across all possible values of ν. The tests proposed below generally allow for this.

As stated previously, we specifically focus on situations where V is ordinal and where the measurement invariance violation is related to the ordinal variable (e.g., the violation is of the type from (5) or the violation grows/shrinks with V). Researchers typically test for measurement invariance w.r.t. ordinal V by employing the alternative from (4), which implicitly treats V as categorical. Thus, test statistics that explicitly treat V as ordinal should have higher power to detect measurement invariance violations that are monotonic

with V.

In the section below, we review the theory underlying tests where V is continuous (and ν is unknown). We then propose novel tests for ordinal V.

Theoretical Detail

This section contains background on the theory underlying the proposed statistics; for a more detailed account, see Merkle and Zeileis (2013).

Model Estimation

We focus specifically on applications where the density $f(\boldsymbol{x}_i; \boldsymbol{\theta})$ arises from a structural equation model with assumed multivariate normality, though the proposed tests extend beyond this family of models. Under the usual regularity conditions (e.g., Ferguson, 1996), the model parameters $\boldsymbol{\theta}$ can be estimated by maximum likelihood (ML), i.e.,

$$\hat{\boldsymbol{\theta}} \;=\; \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}; x_1, \ldots, x_n), \tag{6}$$

or equivalently by solving the first order conditions

$$\sum_{i=1}^{n} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_i) \;=\; 0, \tag{7}$$

where

$$\boldsymbol{s}(\boldsymbol{\theta}; x_i) \;=\; \left(\frac{\partial \ell(\boldsymbol{\theta}; x_i)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}; x_i)}{\partial \theta_k} \right)^\top, \tag{8}$$

is the score function of the model (the partial derivative of the casewise likelihood contributions w.r.t. the parameters $\boldsymbol{\theta}$). Evaluation of the score function at $\hat{\boldsymbol{\theta}}$ for $i = 1, \ldots, n$ measures the extent to which the model maximizes each individual's likelihood: as an individual's scores stray further from zero, the model provides a poorer description of that individual.

Tests for Continuous V

As mentioned previously, when V is categorical with a relatively small number of categories, tests of measurement invariance typically proceed via multiple-group models. In this situation, we use likelihood ratio tests to compare a model whose parameters differ across groups to a model whose parameters are constrained to be equal across groups. When V is continuous, however, multiple-group models usually cannot be used because there are no existing groups. Instead, we can fit a model whose parameters are restricted to be equal across all individuals and then examine how individuals' scores $\boldsymbol{s}(\hat{\boldsymbol{\theta}}; x_i)$ fluctuate with their values of V. If measurement invariance holds with respect to V, then the scores should randomly fluctuate around zero. Conversely, if measurement invariance does not hold, then the scores should systematically depart from zero. These ideas are related to those underlying the Lagrange multiplier test and are discussed in detail by Merkle and Zeileis (2013). Additionally, these ideas are related to those underlying

modification indexes (e.g., Sörbom, 1989): the modification index is equivalent to a Lagrange multiplier test, and the Lagrange multiplier test is contained in the family described by Merkle and Zeileis (2013). Here, we focus on the tests' properties that are relevant for extending them to the ordinal case.

To formalize the ideas discussed in the previous paragraph, we assume that the observations are ordered w.r.t. V, with $x_{(i)}$ reflecting the data for the individual who has the i^{th}-smallest value of V. We then define the k-dimensional *cumulative score process* as

$$\boldsymbol{B}(t;\hat{\boldsymbol{\theta}}) \;=\; \hat{\boldsymbol{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; x_{(i)}) \qquad (0 \leq t \leq 1) \tag{9}$$

where $\lfloor nt \rfloor$ is the integer part of nt and $\hat{I}$ is some consistent estimate of the covariance matrix of the scores. Natural choices for $\hat{I}$ include the information matrix (which we use in our applications and simulations below) or alternatively some kind of outer product of the scores or sandwich estimator to guard the inference against potential misspecification of the model (see Huber, 1967 for the theoretical foundation and Zeileis, 2006b for a computational framework). Equation (9) simultaneously accounts for the ordering of individuals w.r.t. V and decorrelates the scores associated with each of the k model parameters (which allows us to potentially make inferences separately for each individual model parameter). Using ideas similar to those that were outlined in the previous paragraph, the cumulative score process associated with each model parameter should randomly fluctuate around zero under measurement invariance. Further, there exists a functional central limit theorem that allows us to make formal inference with this cumulative score process. Assuming that individuals are independent and the usual ML regularity conditions hold, it is possible to show that (Hjort & Koning, 2002)

$$\boldsymbol{B}(\cdot;\hat{\boldsymbol{\theta}}) \;\xrightarrow{d}\; \boldsymbol{B}^0(\cdot), \tag{10}$$

where $\xrightarrow{d}$ denotes convergence in distribution and $\boldsymbol{B}^0(\cdot)$ is a k-dimensional Brownian bridge. Thus, we can construct tests of measurement invariance by comparing the behavior of the cumulative score process to that of a Brownian bridge. This is accomplished by comparing a scalar statistic associated with the cumulative score process to the analogous statistic of a Brownian bridge.

In practice, we have a finite sample size n and so the *empirical* cumulative score can be represented within an $n \times k$ matrix with elements $\boldsymbol{B}(i/n;\hat{\boldsymbol{\theta}})_j$ that we also denote $\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}$ below for brevity. Each row of the matrix contains cumulative sums of the scores of individuals who were at the i/n percentile of V or below. Scalar test statistics are then obtained by collapsing over rows (individuals) and columns (parameters) of the matrix, with asymptotic distributions of the test statistics under (3) being obtained by applying the same functional to the Brownian bridge (Hjort & Koning, 2002; Zeileis & Hornik, 2007).

Specific test statistics commonly obtained under this framework include the double maximum statistic

$$DM = \max_{i=1,\ldots,n} \max_{j=1,\ldots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}|, \tag{11}$$

which essentially tests whether any component of the cumulative score process strays too

far from zero and is easily visualized. This test discards information related to multiple parameters fluctuating simultaneously, resulting in it having relatively low power for assessing measurement invariance when multiple factor analysis parameters change simultaneously (Merkle & Zeileis, 2013).

Test statistics that exhibit better performance in such situations aggregate information across parameters and possibly also across individuals. These test statistics include

$$CvM = n^{-1} \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \tag{12}$$

$$\max LM = \max_{i=\underline{i},\ldots,\bar{i}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \tag{13}$$

with the former being a Cramér-von Mises statistic and the latter corresponding to a "maximum" Lagrange multiplier test, where the maximum is taken across all possible divisions of individuals into two groups w.r.t. V. Additionally, the $\max LM$ statistic is scaled by the asymptotic variance $t(1-t)$ of the process $\boldsymbol{B}(t, \hat{\boldsymbol{\theta}})$. In simulations, Merkle and Zeileis (2013) found that both tests perform well when assessing simultaneous changes in multiple factor analysis parameters, with the *CvM* test being somewhat advantageous in their particular simulation setup. These simulations included situations in which subsets of model parameters were tested; such situations are handled by focusing only on those columns of $\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}$ that correspond to the parameters of interest.

Proposed Tests for Ordinal V

The theory described above was designed for situations where V is continuous, so that there is a unique ordering of individuals with respect to V. However, in situations where V is ordinal, there is only a partial ordering of all individuals, i.e., observations with the same level of V have no unique ordering. (Note that the same also applies if V is continuous in nature but is only discretely measured, leading to many ties.)

The ordinal statistics proposed here are similar to those described in Equations (11) and (13) above, except that we focus on "bins" of individuals at each level of the ordinal variable. That is, instead of aggregating over all $i = 1, \ldots, n$ individuals, we first compute cumulative proportions t_ℓ ($\ell = 1, \ldots, m-1$) associated with the first $m-1$ levels of V. We then aggregate the cumulative scores only over $i_\ell = \lfloor n \cdot t_\ell \rfloor$. Test statistics related to (11) and (13) above can then be written as

$$WDM_o = \max_{i \in \{i_1,\ldots,i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1/2} \max_{j=1,\ldots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}|, \tag{14}$$

$$\max LM_o = \max_{i \in \{i_1,\ldots,i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \tag{15}$$

resulting in a "weighted" double maximum statistic (weighted by the asymptotic variance of the Brownian bridge) and an ordinal, maximum Lagrange multiplier statistic. Critical values associated with these test statistics can be obtained by applying the same

functionals to bins of a Brownian bridge, where the bin sizes result in the cumulative proportions t_ℓ ($\ell = 1, \ldots, m-1$) associated with the observed V.

For the WDM_o statistic, the resulting asymptotic distribution is $\max_{j=1,\ldots,k} \max_{\ell=1,\ldots,m-1} \boldsymbol{B}^0(t_\ell)/\sqrt{t_\ell(1-t_\ell)}$. Note that the effect of the outer maximum can be easily captured by a Bonferroni correction, as the k components of the Brownian bridge are asymptotically independent. Moreover, the inner maximum is taken over $m-1$ variables $\boldsymbol{B}^0(t_\ell)/\sqrt{t_\ell(1-t_\ell)}$ which are standard normal (due to the scaling with the standard deviation of a Brownian bridge) and have a simple correlation structure: $\sqrt{s(1-t)}/\sqrt{t(1-s)}$ for $s \leq t$ and both $\in \{t_1, \ldots, t_{m-1}\}$. Therefore, critical values and p-values can be easily computed from a multivariate normal distribution with standard normal marginals and this particular correlation matrix; see also Hothorn and Zeileis (2008) for more details. In R, this can be accomplished using the mvtnorm package (Genz et al., 2012).

For $\max LM_o$ the resulting asymptotic distribution is $\max_{\ell=1,\ldots,m-1} ||\boldsymbol{B}^0(t_\ell)||_2^2/(t_\ell(1-t_\ell))$ for which no simple closed-form solution is available. However, critical values and p-values can be obtained through repeated simulation of Brownian bridges. This functionality is built in to R's strucchange package (Zeileis, 2006a), which can be used to generally carry out the tests. Note that for models with only a single parameter to be tested (i.e., $k = 1$) both test statistics are equivalent because then $\max LM_o = WDM_o^2$.

If V is only nominal/categorical, there is not even a partial ordering, i.e., measurement invariance tests should neither exploit the ordering of V's levels nor of the observations within the level. In this situation, it is possible to obtain a test statistic by first summing scores within each of the m levels of the auxiliary variable, then "summing the sums" to obtain a test statistic (Hjort & Koning, 2002). This test statistic can be formally written as

$$LM_{uo} = \sum_{\ell=1,\ldots,m} \sum_{j=1,\ldots,k} \left(\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_\ell j} - \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_{\ell-1} j} \right)^2, \tag{16}$$

where the first subscript on the two terms in parentheses are i_ℓ and $i_{\ell-1}$, respectively (and where we take $i_0 = 0$, so that the cumulative score is $\boldsymbol{B}(0, \hat{\boldsymbol{\theta}}) = \boldsymbol{0}$). Again, tests of subsets of model parameters can be obtained by taking the inner sum over only the $k^* < k$ parameters of interest. This test statistic discards the ordinal nature of the auxiliary variable, essentially employing the alternative hypothesis from (4). A similar issue is observed in testing for measurement invariance via multiple groups models and likelihood ratio tests (or, equivalently, via Wald tests or Lagrange multiplier tests): we can allow $\boldsymbol{\theta}$ to be unique at each level of the ordinal variable, but the ordinality of the auxiliary variable is lost. In contrast, the statistics proposed above explicitly account for the fact that V is ordinal.

As demonstrated in the simulations below, the proposed ordinal test statistics are sensitive to the measurement invariance violations that an analyst would typically expect from an ordinal V. In particular, due to computing cumulative sums in $\boldsymbol{B}(\hat{\boldsymbol{\theta}})$, violations that occur as we move along the levels of V can be captured well. This includes abrupt

shifts in the parameters $\boldsymbol{\theta}$ at a certain level of V as well as smooth increases/decreases in the parameters. Taking a maximum over the k parameters as in WDM_o will be more sensitive to changes that occur only in one out of many parameters, while $\max LM_o$ will be more sensitive to changes occurring in several (or even all of the) parameters simultaneously. Moreover, the test statistics are rather insensitive to anomalies in a small number of categories of V that are unrelated to the ordering of V. This is especially relevant to situations in which the analyst has a large sample size, so that the usual likelihood ratio test is overly sensitive to minor parameter instabilities (e.g., Bentler & Bonett, 1980).

Simulation 1: Detecting Ordinal Invariance Violations

In this simulation, we demonstrate that the proposed test statistics are sensitive to ordinal measurement invariance violations, moreso than traditional statistics. We generate data from a two-factor, six-indicator model, with a measurement invariance violation occurring in the unique variance parameters. We use the proposed test statistics to test for measurement invariance simultaneously across the unique variances, which is similar to a test of "invariant uniquenesses" (see Vandenberg & Lance, 2000).

The two "traditional" statistics that we consider generally treat the ordinal auxiliary variable as categorical. These include the likelihood ratio test of measurement invariance in the six unique variance parameters and the unordered LM test from (16). At the request of reviewers, we also considered the Satorra-Bentler (2001) scaled likelihood ratio test statistic with correction for difference testing, the Yuan-Bentler (1997) scaled test statistic, and the use of AIC (Akaike, 1974) for detecting measurement invariance. We do not report the latter results, because these test statistics performed worse than the usual likelihood ratio test (both here and in Simulation 2).

Method

Data were generated from a two-factor model lacking measurement invariance in the six unique variance parameters. Magnitude of measurement invariance violation, sample size, and number of categories of the ordinal variable were manipulated. We examined three sample sizes ($n = 120, 480, 960$), three numbers of categories ($m = 4, 8, 12$), and seven magnitudes of invariance violations. The measurement invariance violations began at level $1 + m/2$ of V and were constant thereafter. The unique variances for the "violating" levels deviated from the lower levels' unique variances by d times the parameters' asymptotic standard errors (scaled by $\sqrt{n}$), with $d = 0, 0.25, 0.5, \ldots, 1.5$.

For each combination of $n \times m \times d$, 5,000 datasets were generated and tested via the 4 statistics described above. In all conditions, we maintained equal sample sizes at each level of the ordinal variable (i.e., $t_\ell = \ell/m$).

Results

Simulation results comparing the ordinal tests to the unordered LM test and the LRT are presented in Figure 1. Rows of the figure correspond to n, columns of the figure correspond to m, the x-axis of each panel corresponds to d, and the y-axis of each panel

corresponds to power. It is seen that one of the proposed test statistics, the max LM_o statistic from (15), generally has the largest power to detect the ordinal measurement invariance violations. The other three tests are considerably closer in power, with the second proposed ordinal statistic (the double-max test from (14)) exhibiting the lowest power at large violation magnitudes. This is because the double-max test discards information about multiple parameters changing together at specific levels of the ordinal variable (see Merkle & Zeileis, 2013, for related discussion), while the three other tests under consideration make use of this information. Finally, it is seen that, in the small n and large m conditions, the likelihood ratio test exhibits large Type-I error rates (i.e., power greater than 0.05 at $d = 0$). This is because the likelihood ratio test requires estimation of a multiple-groups model, which is very unstable with large numbers of groups and small sample sizes (as only n/m observations are available in each subsample). The statistics proposed here are all of the LM-type and just require estimation of the single-group model, leading to a clear advantage in these conditions.

To summarize, we found the max LM_o statistic to be advantageous for detecting measurement invariance violations that are related to an ordinal auxiliary variable. In particular, power is generally higher, and the test does not require estimation of a multiple group model. Thus, the statistic allows reasonable measurement invariance tests to be carried out at small n/large m combinations. To further illustrate that the proposed statistics are useful for testing violations related to an ordinal variable, we now compare their performance to the likelihood ratio test at large n and small d.

Simulation 2: Minor Anomalies and Large n

In this simulation, we demonstrate that the proposed statistics are relatively insensitive to minor parameter violations that are unrelated to the ordering of the auxiliary variable. As noted earlier, this feature is especially applicable to situations where one's sample size is very large. Analysts often resort to informal fit measures in practice, because the traditional LRT is nearly guaranteed to result in significance. This simulation is intended to show that the proposed ordinal tests remain viable for large n.

Method

Data were generated from the same factor analysis model that was used in Simulation 1, with measurement invariance violations in the unique variance parameters. To implement a minor measurement invariance violation, the unique variances were equal across all levels of the ordinal variable except one (level $1 + m/2$). At this particular level, the unique variances were greater by a factor of d times the parameters' asymptotic standard errors (scaled by $\sqrt{n}$), with $d = 0, 0.5, 1.0, \ldots, 3.0$. The number of levels of the ordinal variable were the same as those in Simulation 1 ($m = 4, 8, 12$), and sample sizes were set at $n = 1200, 4800, 9600$. All other simulation features match those from Simulation 1.

Results

Simulation results for the two ordinal test statistics, the unordered LM test, and the LRT are presented in Figure 2. It is observed that results are very consistent across the

sample sizes tested, implying that "practical infinity" is reached for this model by $n = 1200$. We also observe a negative relationship between power and m; this is because the measurement invariance violation occurred at only one level of the auxiliary variable. As m increases (and n is held constant), the number of individuals violating measurement invariance therefore decreases. As a result, power to detect the violation decreases with increasing m.

The more interesting result of Figure 2 lies in the comparison of the four test statistics within each panel of the figure. The two "unordered" test statistics both have relatively high power to detect the measurement invariance violation, illustrating the result of Bentler and Bonett (1980) and others that the likelihood ratio test statistic picks out minor parameter discrepancies at large n. In contrast, the two ordinal test statistics that we proposed have considerably lower power, with the WDM_o statistic being the lowest and the max LM_o statistic being higher at larger values of d.

These results demonstrate that the proposed ordinal test statistics can be especially useful at large sample sizes, where traditional test statistics result in frequent significance. Both statistics exhibited much lower power to detect a measurement invariance violation that occurs only at a single level of V.

Taken together, the results from Simulation 1 and Simulation 2 provide evidence that the max LM_o statistic should be preferred to the WDM_o statistic for simultaneously assessing measurement invariance across multiple parameters in factor analysis models. The max LM_o statistic has higher power to detect ordinal violations, and its power to detect non-ordinal violations was similar to that of WDM_o when the violation magnitude was small (e.g., for $d \leq 1.5$). The max LM_o statistic is advantageous because it can make use of invariance violations that simultaneously occur in multiple parameters, whereas the WDM_o focuses only on the parameter with the largest invariance violation. Thus, the statistics are likely to exhibit similar performance if only a single model parameter violated measurement invariance. The only disadvantage to max LM_o is that its critical- and/or p-values must be computed by simulation, which can significantly increase computation time. We return to this issue in the General Discussion.

In the next section, we compare the proposed statistics to the likelihood ratio test with real data.

Application: Youth Gratitude

Background

With the positive psychology movement, the construct of gratitude has received much research attention (for a review, see Emmons & McCullough, 2004). Recently, researchers have begun to explore gratitude in youth. One potential problem with this is that researchers, with no exception, have used adult gratitude inventories to measure youth gratitude, thus raising the question of whether the existing gratitude scales used with adults are valid in research with youth. Addressing this issue, Froh et al. (2011) had a large sample of youth ($n = 1401$, ranging from late childhood (10 years old) to late adolescent (19 years old)) complete the three most widely used adult gratitude inventories, including Gratitude Questionnaire 6 (GQ-6; McCullough, Emmons, & Tsang, 2002), Gratitude Adjective Checklist (GAC; McCullough et al., 2002), and Gratitude,

Resentment, Appreciation Test-Short Form (GRAT-Short Form; Thomas & Watkins, 2003). The authors were interested in whether the youth factor structure for the gratitude scales resembles that of adults, and whether the gratitude scales are invariant across the youth age groups.

Method

Froh et al. (2011) used confirmatory factor models to study the invariance of three youth gratitude scales across students aged 10 to 19 years. Due to sample size constraints, the age variable included six categories: 10–11 years, 12–13 years, 14 years, 15 years, 16 years, and 17–19 years. Thus, age is an ordinal variable to which the proposed tests can be applied.

To test for measurement invariance w.r.t. age, each of the three scales was individually factor-analyzed using the items that comprised the scale. For each model, the authors first fit a congeneric model (all parameters free for each level of age), followed by a tau-equivalent model (factor loadings restricted to be equal across each level of age) and a parallel model (all parameters restricted to be equal across levels of age). Because their sample size was large ($n \approx 1400$), they could not rely solely on likelihood ratio tests (i.e., χ^2 difference tests) for model comparison because the tests were overly sensitive at their sample size. To supplement these tests, Froh et al. (2011) examined a set of alternative fit indices, including the non-normed fit index, the comparative fit index, and the incremental fit index (e.g., Browne & Cudeck, 1993). The authors generally found support for the tau-equivalent models through these alternative fit indices: the likelihood ratio test often resulted in significance even when the alternative indices indicated good fit.

In the analyses described below, we re-analyze the Froh et al. (2011) data using the ordinal test statistics proposed in this paper. This results in a series of tests that are less sensitive than the likelihood ratio test to minor parameter discrepancies, while being more sensitive to ordinal violations of measurement invariance. We focus on two analyses from Froh et al. (2011) where the likelihood ratio test resulted in significance (indicating that the restricted model did not fit as well as the less-restricted model) but the alternative fit measures indicated the opposite. These include comparison of a one-factor congeneric model to a one-factor tau-equivalent model using the GQ6 and comparison of a one-factor tau-equivalent model to a one-factor parallel model using the GAC. To conduct equivalent analyses via the proposed tests, we fit the more-restricted model in each case and test for instability in the focal model parameters.

Of the 1401 cases originally collected by Froh et al. (2011), we use here all subjects with complete data (resulting in $n = 1327$).

Results

The results section is divided into two subsections, one for each analysis described above. The first subsection contains an example of the ordinal statistics disagreeing with the likelihood ratio tests, while the second subsection contains the opposite.

GQ-6. In fitting a tau-equivalent model and a congeneric model to the GQ-6 data, Froh et al. (2011) used alternative fit indices to conclude that the tau-equivalent model was

as good as the congeneric model. However, the likelihood ratio test comparing these two models was significant ($\chi^2_{20} = 38.08, p = 0.009$ for the data considered here).

We can use the proposed ordinal statistics to assess whether or not the factor loadings in the tau-equivalent model fluctuate with respect to age. Unlike the LRT, the test does not require parameters to differ across *all* subgroups. Instead, we test for deviations such that a split into two subgroups is sufficient to capture the effect of V. In employing the ordinal tests, we obtain $WDM_o = 2.91, p = 0.060$ and $\max LM_o = 11.16, p = 0.096$. Both p-values are clearly larger than that of the likelihood ratio test and neither is significant at $\alpha = 0.05$, which supports the conclusions that Froh et al. (2011) obtained from alternative fit statistics. This provides further evidence that there is no systematic deviation of the factor loadings along age and that the likelihood ratio statistic is overly sensitive, picking up some non-systematic dependence on age.

Plots representing the statistics' fluctuations across levels of age group are displayed in Figure 3. The left panel displays the process associated with WDM_o from (14), i.e., the sequence of weighted maximum (over j) statistics for each potential threshold i. The right panel displays the process associated with $\max LM_o$ from (15), i.e., the sequence of LM statistics for each potential threshold i. In both cases, the test statistics in the sequence assess a split of the observations up to age group i vs. greater than i, and the null hypothesis is rejected if the maximum of the statistics is larger than its 5% critical value (visualized by the horizontal red line). Therefore, the final age group (17–19 years) is not displayed, because the statistics associated with this final age group would encompass all observations in a single group and hence always equal zero. It is observed that both statistics generally increase with age, with WDM_o being largest for a threshold of 15 years and $\max LM_o$ for a threshold of 16 years. The differing pattern of values for the 15- and 16-year-olds can be taken as an indication that some factor loading is unstable at an age of 15 or 16, but this is not a clear and general trend across all the tested loadings and age groups.

GAC. In fitting tau-equivalent and parallel models to the GAC data, Froh et al. (2011) obtained mixed results. The alternative fit indices did not all agree with one another, and the likelihood ratio test indicated that the parallel model fit worse than the tau-equivalent model ($\chi^2_{20} = 167.72, p < 0.01$ for the data considered here). Froh et al. (2011) ultimately concluded that the tau-equivalent model provided a better fit than did the parallel model.

To apply the ordinal statistics proposed in this paper, we fit the parallel model and test for instability in the variance parameters (unique variance and factor variance) w.r.t. age. This results in $WDM_o = 6.55, p < 0.01$ and $\max LM_o = 113.13, p < 0.01$. Both of these statistics agree with the general conclusion that the parallel model is not sufficient, providing further evidence that the significant likelihood ratio test is not simply an artifact of the large sample size.

Plots representing the statistics' fluctuations across age groups are displayed in Figure 4. The left panel displays the process associated with WDM_o, while the right panel displays the process associated with $\max LM_o$. It is observed that both processes are fully above the critical value, implying the measurement invariance violation. Additionally, both processes peak at the 12–13 age group, suggesting that parameters differ between individuals up to 13 years of age and individuals older than 13 years of age.

The finding that variance parameters differ between individuals up to 13 years and individuals over 13 years is reinforced by comparing the tau-equivalent and parallel models to an intermediate model. This intermediate model is tau-equivalent in nature, but there exist only two groups: individuals up to 13 years and individuals older than 13 years. A likelihood ratio test implies that this intermediate model fits as well as the original tau-equivalent model ($\chi^2_{16} = 13.72$, $p = 0.62$), with 16 fewer parameters ($= 6 \cdot 4 - 2 \cdot 4$ because the four variances have to be estimated in only two rather than six age groups). The intermediate model also fits better than the parallel model, as judged by a second likelihood ratio test ($\chi^2_4 = 154.00, p < 0.01$). Finally, using the proposed ordinal test statistics with the intermediate model, we no longer observe further instability in the variance parameters ($WDM_o = 1.84, p = 0.86$; $\max LM_o = 3.83, p = 0.99$).

Summary

The application considered above shows that the ordinal test statistics can provide useful information in situations where one might question significant likelihood ratio test statistics. While researchers use rules of thumb to obtain decisions from other alternative fit measures, the proposed statistics are proper tests of the hypothesis of interest. They can be used to either supplement or replace the likelihood ratio test, depending upon the types of measurement invariance violations in which the researcher has a priori interest. We further describe the issue of supplementing vs. replacing the likelihood ratio test in the general discussion.

General Discussion

In this paper, we proposed two statistics that can be used when one has an ordinal auxiliary variable and wishes to study measurement invariance. We demonstrated via simulation that these statistics have good properties, though these results necessarily examined a small number of models and invariance violations and may not hold in all situations. To our knowledge, the ordinal measurement invariance statistics proposed here are the only ones that treat auxiliary variables as ordinal and thus direct power against alternatives that are typically of interest to practitioners. Other methods treat the auxiliary variable as either continuous or categorical, in a manner similar to the treatment of ordinal predictor variables in linear regression. In the remainder of the paper, we provide detail on test choice and on the tests' applicability to other models.

Choice of Test

The results presented in this paper imply that the proposed ordinal statistics may "miss" measurement invariance violations that are not monotonic w.r.t. V. More precisely, while the suggested tests are also consistent for such non-monotonic violations, they seem to be less powerful than the likelihood ratio test. We speculate that, in most applications, this will not be a major issue because the researcher's a priori hypotheses exclusively focus on monotonic measurement invariance violations. For example, in the youth gratitude application, we tested for measurement invariance across six age groups. If we observed a measurement invariance violation whereby factor loadings were equal at all age groups

except 14 years, we would have a hard time explaining the violation as anything but an anomaly in the 14-year-olds. Further, if n is large, we are likely to suspect that the result arises from the large sample size. We may still be interested in why the 14-year-olds differed, but the analysis is purely exploratory at this point because this type of violation was unexpected. However, there is generally a tradeoff between the ordinal statistics and the likelihood ratio statistic. The proposed ordinal statistics usually provide more powerful tests of one's a priori hypothesis regarding measurement invariance w.r.t. ordinal V, while the likelihood ratio statistic provides a more powerful test of general (non-monotonic) measurement invariance w.r.t. V. While the latter feature may be important in some high-stakes applications, many researchers are likely to find the former feature appealing for their work.

Along with using likelihood ratio tests to study measurement invariance, researchers may wish to treat ordinal V as continuous (especially if V has very many levels). As described in detail by Merkle and Zeileis (2013), we can also use cumulative score processes with continuous V, resulting in, e.g., maximum LM statistics and Cramér-von Mises statistics. In fact, when the number of potential thresholds is large, the proposed $\max LM_o$ statistic will be very close to the $\max LM$ statistic described in Merkle and Zeileis (2013). Thus, the formation of ordinal age groups (or other variables) is not necessary for testing measurement invariance, and it may be beneficial to collect continuous age data (e.g., age measured in days rather than in years).

There also exist alternative methods for testing measurement invariance w.r.t. continuous V, including moderated factor models (Bauer & Hussong, 2009; Molenaar, Dolan, Wicherts, & van der Mass, 2010; Purcell, 2002) and factor mixture models (Dolan & van der Maas, 1998; Lubke & Muthén, 2005). Under the moderated factor model approach, V is inserted directly into the factor analysis model and allowed to have a linear relationship with model parameters. Under the factor mixture model approach, individuals are typically assumed to arise from a small number of distinct factor analysis models. The ordinal variable V could then be used to predict the probability that an individual arises from each model. These treatments of ordinal V as continuous will often be advantageous, especially if the levels of V are approximately equally-spaced and the relationship between V and the measurement invariance violation is linear. The approaches do require models of greater complexity and may not be suitable for all ordinal V, however, while the methods we propose here are generally suitable for ordinal V.

Finally, a practical issue associated with the proposed $\max LM_o$ statistic involves the computation of p-values: p-values and/or critical values must be computed via simulation of a Brownian bridge, with the simulation depending on the relative proportion of cases at each level of the ordinal auxiliary variable. Hence, a new simulation usually must be conducted for each dataset, which can be somewhat time-consuming (on the order of minutes, as opposed to seconds or hours).

Extension to Other Models

We focused on testing for measurement invariance in factor analysis models here, but the proposed test statistics are applicable to other psychometric models that are estimated via ML (or similar estimation techniques for independent observations that are governed by

a central limit theorem). The only requirement for carrying out the tests is that the casewise scores (Equation (8)) be available following model estimation. As a result, applications to studying DIF in IRT are immediate (for a presentation involving non-ordinal variables, see Strobl et al., 2013), as are general psychometric applications to studying parameter stability w.r.t. ordinal auxiliary variables. These could include applications where ordinal variables are explicitly included in the model, such as ordinal factor analysis. We expect the same general results to hold for these applications, whereby the proposed test statistics are better than the LRT for detecting monotonic instabilities. The strucchange package (Zeileis, 2006a) noted previously can be used for these more-general applications.

Summary

As demonstrated via simulation, the proposed test statistics have relatively high power for detecting measurement invariance violations that are monotonic with the ordinal variable, and they have relatively low power for detecting minor violations that are not monotonic. The former feature implies that the statistics are good at detecting measurement invariance violations that are interpretable to the researcher, while the latter feature implies that the statistics are feasible in situations where the likelihood ratio test commonly rejects H_0 in practice (e.g., Bentler & Bonett, 1980). Furthermore, the focal psychometric model does not have to be modified in any way, which differs from approaches that may treat the ordinal variable as continuous. In all, the tests have advantageous properties that should be useful in practice.

Computational Details

All results were obtained using the R system for statistical computing (R Development Core Team, 2012), version 3.0.2, employing the add-on package lavaan 0.5-14 (Rosseel, 2012) for fitting of the factor analysis models and strucchange 1.5-0 (Zeileis, Leisch, Hornik, & Kleiber, 2002; Zeileis, 2006a) for evaluating the parameter instability tests. R and the packages lavaan and strucchange are freely available under the General Public License 2 from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`. R code for replication of our results is available at `http://semtools.R-Forge.R-project.org/`.

References

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

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*, 101–125.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park: Sage Publications.

Dolan, C. V., & van der Maas, H. L. J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*, 227–253.

Emmons, R. A., & McCullough, M. E. (Eds.). (2004). *The psychology of gratitude*. New York: Oxford University Press.

Ferguson, T. S. (1996). *A course in large sample theory*. London: Chapman & Hall.

Froh, J. J., Fan, J., Emmons, R. A., Bono, G., Huebner, E. S., & Watkins, P. (2011). Measuring gratitude in youth: Assessing the psychometric properties of adult gratitude scales in children and adolescents. *Psychological Assessment*, *23*, 311–324.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2012). mvtnorm: Multivariate normal and t distributions [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=mvtnorm` (R package version 0.9-9992)

Hjort, N. L., & Koning, A. (2002). Tests for constancy of model parameters over time. *Nonparametric Statistics*, *14*, 113–132.

Hothorn, T., & Zeileis, A. (2008). Generalized maximally selected statistics. *Biometrics*, *64*(4), 1263–1269.

Huber, P. J. (1967). The behavior of maximum likelihood estimation under nonstandard conditions. In L. M. LeCam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.* Berkeley: University of California Press.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21–39.

McCullough, M. E., Emmons, R. A., & Tsang, J.-A. (2002). The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology*, *82*, 112–127.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*, 59–82.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Mass, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using

moderated factor analysis. *Intelligence*, *38*, 611–624.

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research*, *5*, 554–571.

R Development Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. URL `http://www.R-project.org/`. Vienna, Austria. (ISBN 3-900051-07-0)

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

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

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384.

Strobl, C., Kopf, J., & Zeileis, A. (2013). A new method for detecting differential item functioning in the Rasch model. *Psychometrika*. (Forthcoming)

Thomas, M., & Watkins, P. (2003). *Measuring the grateful trait: Development of the revised GRAT.* Poster presented at the Annual Convention of the Western Psychological Association, Vancouver, BC.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*, 767–774.

Zeileis, A. (2006a). Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics & Data Analysis*, *50*(11), 2987–3008.

Zeileis, A. (2006b). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*(9), 1–16. Retrieved from `http://www.jstatsoft.org/v16/i09/`

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*, 488–508.

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*, 1–38. Retrieved from `http://www.jstatsoft.org/v07/i02/`

Figure 1. Simulated power curves for the ordered and unordered max LM tests, the ordered double-max test, and the likelihood ratio test across three sample sizes n, three levels of the ordinal variable m, and measurement invariance violations of 0–1.5 standard errors (scaled by $\sqrt{n}$).

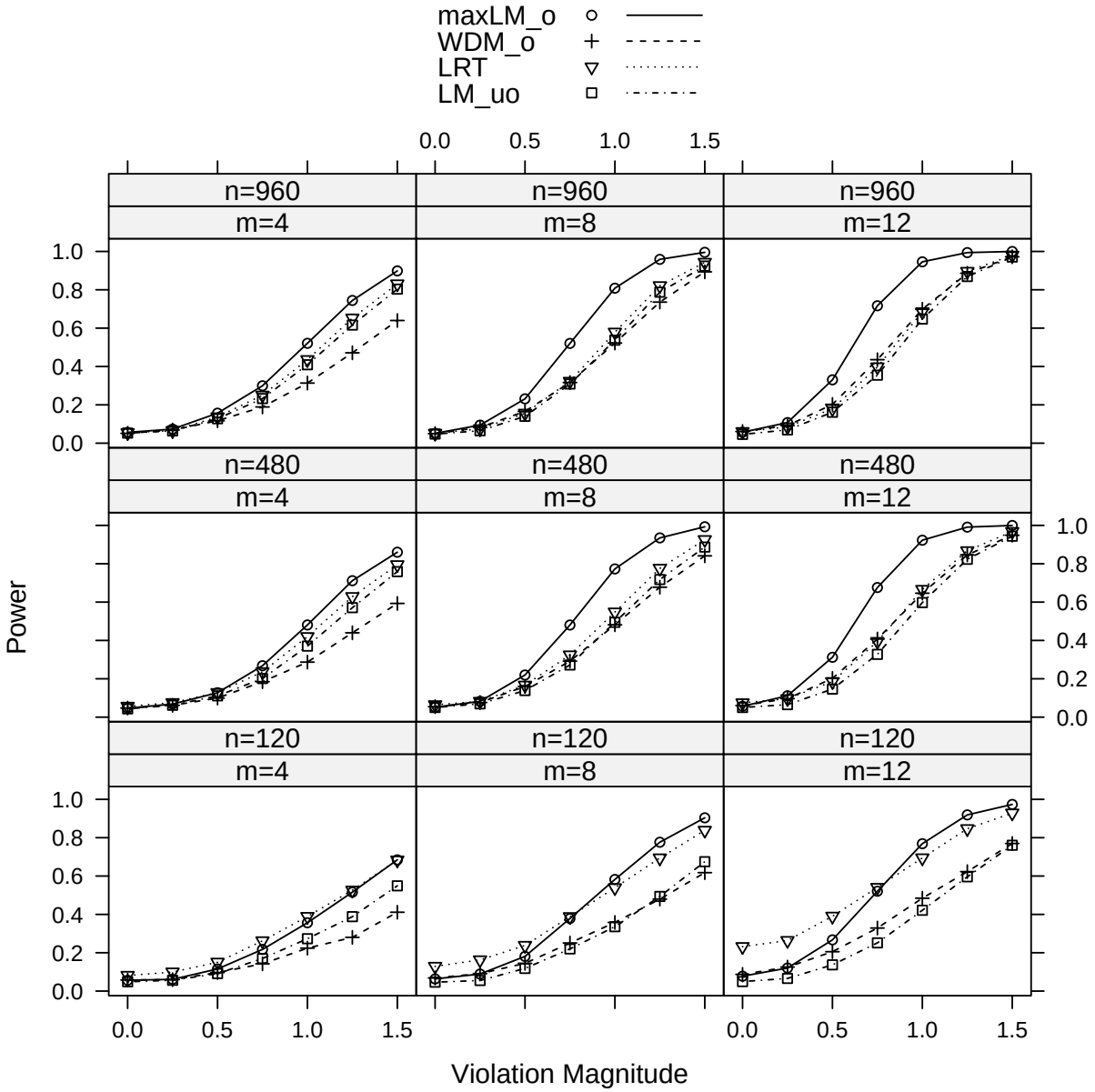

Figure 2. Simulated power curves for the ordered and unordered max LM tests, the ordered double-max test, and the likelihood ratio test across three sample sizes n, three levels of the ordinal variable m, and measurement invariance violations of 0–3 standard errors (scaled by $\sqrt{n}$) occurring at a single level (the $(1 + m/2)^{\text{th}}$ level) of the ordinal auxiliary variable.

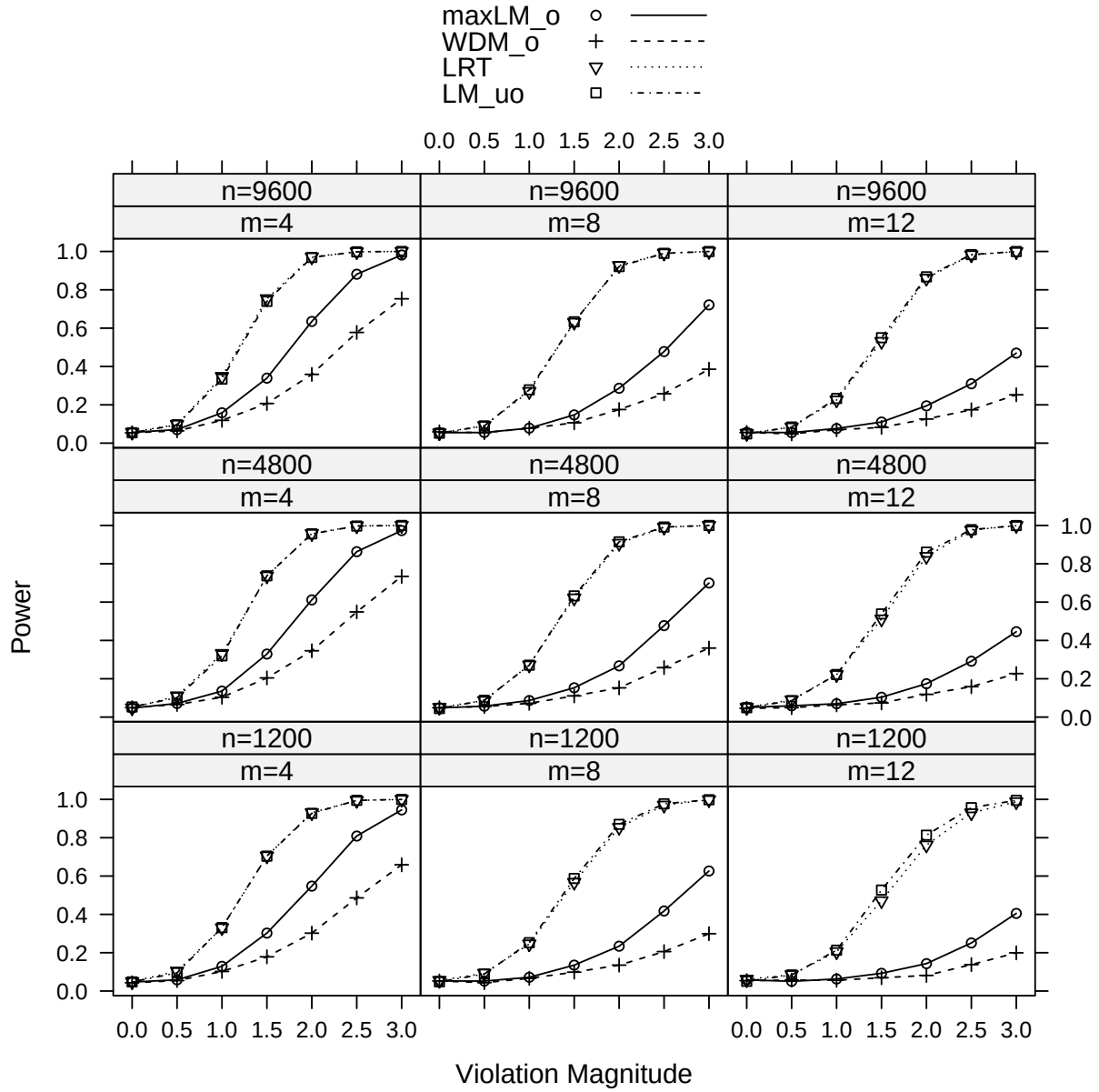

Figure 3. Fluctuation processes for the WDM_o statistic (left panel) and the max LM_o statistic (right panel), arising from the GQ-6 data.

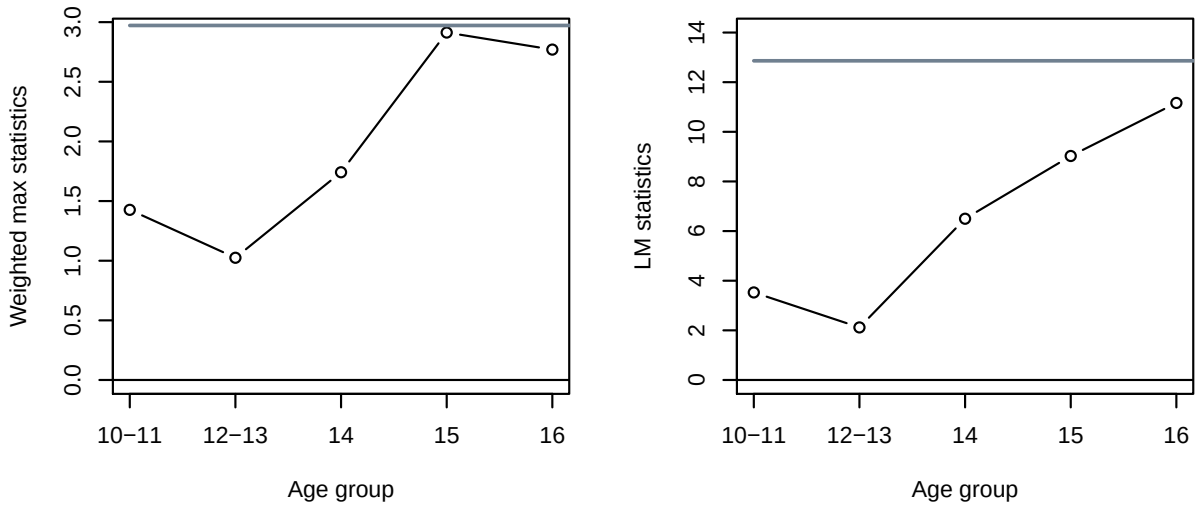

Figure 4. Fluctuation processes for the WDM_o statistic (left panel) and the max LM_o statistic (right panel), arising from the GAC data.