University of Innsbruck

Score-based tests of measurement invariance: Use in practice

Ting Wang, Edgar C. Merkle, Achim Zeileis

Working Papers in Economics and Statistics

2013-33

Research Platform
eeecon
Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact Address:
University of Innsbruck
Department of Public Finance
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 Measurement Invariance: Use in Practice

Ting Wang
University of Missouri

Edgar C. Merkle
University of Missouri

Achim Zeileis
Universität Innsbruck

Abstract

In this paper, we consider a family of recently-proposed measurement invariance tests that are based on the *scores* of a fitted model. This family can be used to test for measurement invariance w.r.t. a continuous auxiliary variable, without pre-specification of subgroups. Moreover, the family can be used when one wishes to test for measurement invariance w.r.t. an ordinal auxiliary variable, yielding test statistics that are sensitive to violations that are monotonically related to the ordinal variable (and less sensitive to non-monotonic violations). The paper is specifically aimed at potential users of the tests who may wish to know (i) how the tests can be employed for their data, and (ii) whether the tests can accurately identify specific models parameters that violate measurement invariance (possibly in the presence of model misspecification). After providing an overview of the tests, we illustrate their general use via the R packages **lavaan** and **strucchange**. We then describe two novel simulations that provide evidence of the tests' practical abilities. As a whole, the paper provides researchers with the tools and knowledge needed to apply these tests to general measurement invariance scenarios.

Keywords: measurement invariance, parameter stability, ordinal variable, factor analysis, structural equation models.

1. Introduction

Recently, there has been increasing interest in the topic of approximate measurement invariance: we know that strict hypotheses of measurement invariance do not hold exactly across different groups, and this should be reflected in corresponding tests of measurement invariance. Under a Bayesian approach, we may utilize the idea of approximate invariance (e.g., Muthén and Asparouhov 2013), whereby across-group equality constraints on parameters are replaced with informative prior distributions. In this paper, we describe an alternative approach: the development of test statistics that are sensitive to invariance violations of interest and insensitive to "anomalous" invariance violations. The test statistics are specifically applicable to situations where one wishes to test for measurement invariance with respect to an ordinal variable, and they are special cases of a family of tests that may be used to study measurement invariance w.r.t. continuous, categorical, and ordinal variables.

The family of tests described in this paper have recently been applied to the study of measurement invariance (Merkle and Zeileis 2013; Merkle, Fan, and Zeileis 2013), though their practical application has been limited to a small set of simulations and datasets. In this paper, we provide a detailed illustration of the tests' use and performance under scenarios likely to

be encountered in practice. In the following sections, we first briefly review the theoretical framework of the proposed tests and provide a short tutorial illustrating the use of the tests in R (R Core Team 2013). Next, we study the tests' performance in simulations that mimic practical research scenarios. Finally, we provide some further discussion on the tests' use in practice.

2. Background

This section contains background and discussion of the proposed statistics as applied to structural equation models (SEMs); for a more detailed account, see Merkle and Zeileis (2013) and Merkle *et al.* (2013). For details on the statistics' application to general statistical models, see Zeileis and Hornik (2007).

As currently implemented for SEM, the statistical tests described in this paper can be applied to models that are estimated via a multivariate normal or Wishart likelihood (or discrepancy) function, with extension to other discrepancy functions being conceptually straightforward. The tests are carried out following model estimation, making use of output associated with the fitted model. In general, we fit a model that restricts parameters to be equal across observations, then carry out a posthoc test to examine whether specific parameters vary across observations. This procedure is similar in spirit to the calculation of modification indices (Bentler 1990) and to Lagrange multiplier tests (Satorra 1989), and, in fact, those statistics can be viewed as special cases of the family described here.

Following model estimation, the tests primarily work on the *scores* of the fitted model; these are defined as

$$s(\theta; x_i) = \left(\frac{\partial \ell(\theta; x_i)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta; x_i)}{\partial \theta_k} \right)^\top, \quad i = 1, \ldots, n, \tag{1}$$

where $\ell(\theta; x_i)$ is the likelihood associated with individual i and θ is a k-dimensional parameter vector. The corresponding maximum likelihood estimate $\hat{\theta}$ solves the first order condition: $\sum_{i=1}^{n} s(\hat{\theta}; x_i) = 0$. The cross-product of these scores forms the "meat" for the calculation of robust (Huber-White) standard errors (e.g., Zeileis 2006b).

To verbally describe the above equation, each individual has k scores describing the extent to which the fitted model describes that particular individual. Scores close to zero indicate a "good" description, and scores far from zero indicate a "bad" description. Each of an individual's k scores represent one model parameter, so that the scores provide specific information about each parameter's fit to each individual.

To use these scores for testing, we order individuals according to an auxiliary variable V (the variable against which we are testing measurement invariance) and look for "trends" in the scores. For example, imagine that we are testing for measurement invariance w.r.t. age. If the manifest variables violate measurement invariance here, then some parameter estimates may be too large for young individuals and too small for old individuals (say). This result would be reflected in the scores, where young individuals' scores may be greater than zero and old individuals' scores less than zero (though the sign of the scores will depend on whether a function is being minimized or maximized). Conversely, if measurement invariance holds, then all individuals' scores will fluctuate randomly around zero.

To formalize these ideas, we define a suitably scaled cumulative sum of the ordered scores.

This may be written as

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

where $\hat{\boldsymbol{I}}$ is an estimate of the information matrix, $\lfloor nt \rfloor$ is the integer part of nt (i.e., a floor operator), and $x_{(i)}$ reflects the individual with the i-th smallest value of the auxiliary variable V. We specifically focus on how the cumulative sum fluctuates as more individuals' scores are added to it, e.g., starting with the youngest and ending with the oldest individual if age is the auxiliary variable of interest. The summation is premultiplied by an estimate of the inverse square root of the information matrix, which serves to decorrelate the fluctuation processes associated with individual model parameters. Thus, the process associated with one parameter is not correlated with the processes associated with other parameters.

Under the hypothesis of measurement invariance, a central limit theorem can be used to show that the fluctuation of the above cumulative sum follows a Brownian bridge. This result allows us to calculate p-values and critical values for test statistics under the hypothesis of measurement invariance. We can obtain test statistics associated with all model parameters and with subsets of model parameters.

Multiple test statistics are available, depending on how one summarizes the behavior of the cumulative sum of scores. For example, one could take the absolute maximum that the cumulative sum attains for any parameter of interest, resulting in a *double max* statistic (the maximum is taken across parameters and individuals). Alternatively, one could sum the (squared) cumulative sum across parameters of interest and take the maximum or the average across individuals, resulting in a *maximum Lagrange multiplier* statistic and *Cramér-von Mises* statistic, respectively (see Merkle and Zeileis 2013, for further discussion). These statistics are given by

$$DM \;=\; \max_{i=1,\ldots,n} \max_{j=1,\ldots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}| \tag{3}$$

$$CvM \;=\; n^{-1} \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \tag{4}$$

$$\max LM \;=\; \max_{i=\underline{i},\ldots,\bar{i}} \left\{ \frac{i}{n}\left(1-\frac{i}{n}\right)\right\}^{-1} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2. \tag{5}$$

Critical values associated with DM can be obtained analytically, while critical values associated with the other statistics can be obtained from direct simulation (Zeileis 2006a) or from more refined techniques (Hansen 1997). This issue should not be important to the user, as critical values are obtained directly from the R implementation described later.

Importantly, the above statistics were derived for situations where individuals are uniquely ordered according to the auxiliary variable. This is not always the case for measurement invariance applications, where the auxiliary variable is often ordinal. To remedy this situation, Merkle *et al.* (2013) extended the framework to situations where one has an ordinal auxiliary variable of interest. Essentially, one allows all individuals with the same value of the auxiliary variable to enter into the cumulative sum at the same time. Analogous test statistics are then computed, with modified critical values being adopted to reflect the change in the statistics' computation.

For an ordinal auxiliary variable with m levels, these modifications are based on t_ℓ ($\ell = 1, \ldots, m-1$), which are the empirical, cumulative proportions of individuals observed at the first $m-1$ levels. The modified statistics are then given by

$$WDM_o = \max_{i \in \{i_1, \ldots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n}\right) \right\}^{-1/2} \max_{j=1,\ldots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}|, \tag{6}$$

$$\max LM_o = \max_{i \in \{i_1, \ldots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n}\right) \right\}^{-1} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \tag{7}$$

where $i_\ell = \lfloor n \cdot t_\ell \rfloor$ ($\ell = 1, \ldots, m-1$). Critical values associated with the WDM_o statistic can be obtained directly from a multivariate normal distribution (see Hothorn and Zeileis 2008), while critical values associated with $\max LM_o$ can be obtained via simulation. This simulation is somewhat computationally intensive and, in practice, takes about 10 minutes on the authors' computers when 50,000 replications are sampled from the approximate asymptotic distribution. However, the wait is often worth it, as Merkle *et al.* (2013) found the performance of the $\max LM_o$ statistic to have more power than the WDM_o statistic and the traditional likelihood ratio test statistic when the measurement invariance violation is monotonic with the ordinal variable.

Finally, if the auxiliary variable V is only nominal/categorical, the cumulative sums of scores can be used to obtain a Lagrange multiplier statistic. This test statistic can be formally written as

$$LM_{uo} = \sum_{\ell=1,\ldots,m} \sum_{j=1,\ldots,k} \left(\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_\ell j} - \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_{\ell-1} j} \right)^2, \tag{8}$$

where $\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_0 j} = 0$ for all j. This statistic is asymptotically equivalent to the usual, likelihood ratio test statistic, and it is advantageous over the likelihood ratio test because it requires estimation of only one model (the restricted model). We make use of this advantage in the simulations, described later.

3. Tutorial

In this section, we demonstrate how the above tests can be carried out in R, using the package **lavaan** (Rosseel 2012) for model estimation and **strucchange** (Zeileis, Leisch, Hornik, and Kleiber 2002; Zeileis 2006a) for testing. We use data from Froh, Fan, Emmons, Bono, Huebner, and Watkins (2011) concerning the applicability of adult gratitude scales to youth, available in the R package *psychotools* (Zeileis, Strobl, and Wickelmaier 2013). The data consist of responses to three adult gratitude scales from $n = 1401$ youth aged 10–19 years. The original authors were specifically interested in whether the scales were measurement invariant across age. Because the sample size at each age was unbalanced, the authors created age groups of approximately equal sample size. In the examples below, we test for measurement invariance across these age groups.

We focus on measurement invariance of the factor loadings associated with the GQ-6 scale, using a one-factor model. While the "age group" variable against which we are testing measurement invariance is best considered ordinal, for demonstration we consider its treatment as categorical, continuous, and ordinal. Each of these treatments is described below in a separate section.

3.1. Categorical treatment

Measurement invariance is most often tested using multiple groups models (see, e.g., van de Schoot, Lugtig, and Hox 2012). This amounts to assuming that our auxiliary variable is categorical (i.e., unordered), which is not true for the age groups in the data. However, we demonstrate the procedure for completeness.

To conduct the analysis, we first load the data and keep only complete cases for simplicity (though the tests can be applied to incomplete data).

```
R> data("YouthGratitude", package = "psychotools")
R> compcases <- apply(YouthGratitude[, 4:28], 1, function(x) all(x %in% 1:9))
R> yg <- YouthGratitude[compcases, ]
```

Next, we fit two models in **lavaan**: a one-factor model where loadings are restricted to be equal across age groups, and a one-factor model where loadings are free across age groups.

```
R> restr <- cfa("f1 = ~gq6_1 + gq6_2 + gq6_3 + gq6_4 + gq6_5",
+    data = yg, group = "agegroup", meanstructure = TRUE,
+    group.equal = "loadings")
R> full <- cfa("f1= ~gq6_1 + gq6_2 + gq6_3 + gq6_4 + gq6_5",
+    data = yg, group = "agegroup", meanstructure = TRUE)
```

Finally, we can get the results of a likelihood ratio test via the `anova()` function, which implies that the GQ-6 violates measurement invariance.

```
R> anova(full, restr)

Chi Square Difference Test

      Df   AIC   BIC Chisq Chisq diff Df diff Pr(>Chisq)
full  30 18947 19414   139
restr 50 18945 19308   177       38.1      20     0.0087
```

To obtain the asymptotically equivalent LM_{uo} (Equation 8), we can use the `sctest()` function from **strucchange**:

```
R> sctest(restr, order.by = yg$agegroup, parm = 1:4, vcov = "info",
+    functional = "LMuo")

        M-fluctuation test

data:  restr
f(efp) = 31.4, p-value = 0.05018
```

This command specifies that we assess the parameter 1–4 of model `restr` after ordering the observations according to `agegroup`. Additionally, the observed information matrix is used as the variance-covariance matrix. Note that the model parameters 1–4 are the factor

loadings supplied by **lavaan**, which can be seen by inspecting `coef(restr)`. This also leads to somewhat smaller test statistics that are very close to being significant at the 5% level.

Because our sample size is large, the likelihood ratio test is known to be sensitive to small measurement invariance violations (Bentler and Bonett 1980). That is, the above tests are sensitive to small measurement invariance violations that are not likely to be of interest to researchers. For example, imagine that the 15-year-olds' parameters are slightly different than the other age groups. The 15-year-olds are in the middle of the age groups, and there is not likely to be any theoretical justification for 15-year-olds differing from every other age group. One solution to this problem would be the Bayesian, approximate invariance methods described in the introduction (Muthén and Asparouhov 2013). Alternatively, we can use the proposed family of score-based statistics to obtain tests that are sensitive to the ordering of age.

3.2. Continuous treatment

If we are interested in measurement invariance violations that are monotonic with the age groups, it is perhaps simplest to treat the age groups as continuous. In doing so, we can use the statistics from Equations 3, 4, and 5. That is, we can fit a model whose parameters are restricted to be equal across all individuals and then examine how individuals' scores $s(\hat{\theta}; x_i)$ fluctuate with their age (where age ties are broken arbitrarily, using the original order of the observations within each age group). This is demonstrated below, with similar code being useful when one is testing for measurement invariance w.r.t. truly continuous variables.

Again, we employ the `sctest()` function to assess parameters 1–4 from the restricted model `restr` after ordering w.r.t. `agegroup`:

```
R> dm <- sctest(restr, order.by = yg$agegroup, parm = 1:4, vcov = "info",
+    functional = "DM")
R> cvm <- sctest(restr, order.by = yg$agegroup, parm = 1:4, vcov = "info",
+    functional = "CvM")
R> maxlm <- sctest(restr, order.by = yg$agegroup, parm = 1:4,
+    vcov = "info", functional = "maxLM")
R> c(dm$p.value, cvm$p.value, maxlm$p.value)

[1] 0.03804 0.11557 0.00414
```

We see that two of the three p-values output at the end of the code are larger than that associated with the LRT (with the CvM statistic being non-significant).

The tests carried out here assume a unique ordering of individuals by age, but this is obviously not the case. To compute the statistics and p-values, the **strucchange** package implicitly employed the (arbitrary) ordering of individuals who are tied on age. If we were to change this ordering, the resulting statistics and p-values would also change, potentially switching significant results to being non-significant and vice versa. Clearly, this is problematic. To accurately account for the multiple observations at the same age level, we must use the ordinal tests from Equations 6 and 7. These are described next.

3.3. Ordinal treatment

The main difference between the ordinal test statistics and their continuous counterparts is

that the ordinal statistics are unchanged when re-ordering individuals within the same age group. To compute the test statistics, we allow the scores of all tied individuals to enter the cumulative sum (Equation 2) simultaneously. This results in modified critical values and test statistics that are sensitive to measurement invariance violations that are monotonic w.r.t. age group.

To carry out the tests, we can rely on the same function that we used for the continuous test statistics. As mentioned previously, calculation of the max LM_o statistic (Equation 7) can be lengthy from the need to simulate critical values (though see the end of this section, which provides a partial speed-up).

```
R> wdmo <- sctest(restr, order.by = yg$agegroup, parm = 1:4, vcov = "info",
+    functional = "WDMo")
R> maxlmo <- sctest(restr, order.by = yg$agegroup, parm = 1:4,
+    vcov = "info", functional = "maxLMo")
R> c(wdmo$p.value, maxlmo$p.value)

[1] 0.0588 0.0970
```

In computing the ordinal test statistics, we obtain $p = 0.059$ and $p = 0.097$, respectively.[1] Both p-values are clearly larger than that of the likelihood ratio test and neither is significant at $\alpha = 0.05$. This provides evidence that there is no measurement invariance violation that is monotonic with age group. Instead, given the large sample size, the likelihood ratio test may be overly sensitive to anomalous, non-monotonic violations at one (or a few) age groups.

In addition to test statistics, "instability plots" can be generated by setting `plot = TRUE` in the `sctest()` calls above. The resulting plots representing the ordinal statistics' fluctuations across levels of age group are displayed in Figure 1. These plots display the loadings' fluctuation across age groups, with the x-axis reflecting age group and the y-axis reflecting test statistic values (larger values reflect more instability). The dashed horizontal lines reflect critical values, so that the hypothesis of measurement invariance is rejected if the sequence of test statistics crosses the critical value. While the measurement invariance tests are non-significant, the plots imply some instability in the older age groups (15 and 16).

Finally, if the user anticipates multiple calculations of the max LM_o statistic for a specific dataset, it is possible to save time by simulating critical values once and re-using them for multiple tests. We can use the `ordL2BB()` function to generate critical values and store them in an object `mLMo`, say. Then, this object can be employed to obtain the test statistic in the usual manner.

```
R> mLMo <- ordL2BB(yg$agegroup)
R> maxlmo <- sctest(restr, order.by = yg$agegroup, parm = 1:4,
+    vcov = "info", functional = mLMo)
```

The `ordL2BB()` command automatically generates critical values for testing 1 to 20 parameters at a time. If only a smaller number of parameters (e.g., only up to 6) is to be tested, some computation time can be saved by setting the `nproc` argument accordingly (e.g., `nproc =`

[1] To replicate both p-values exactly, R's random seed needs to be set by `set.seed(1090)` prior to each `sctest()` call.

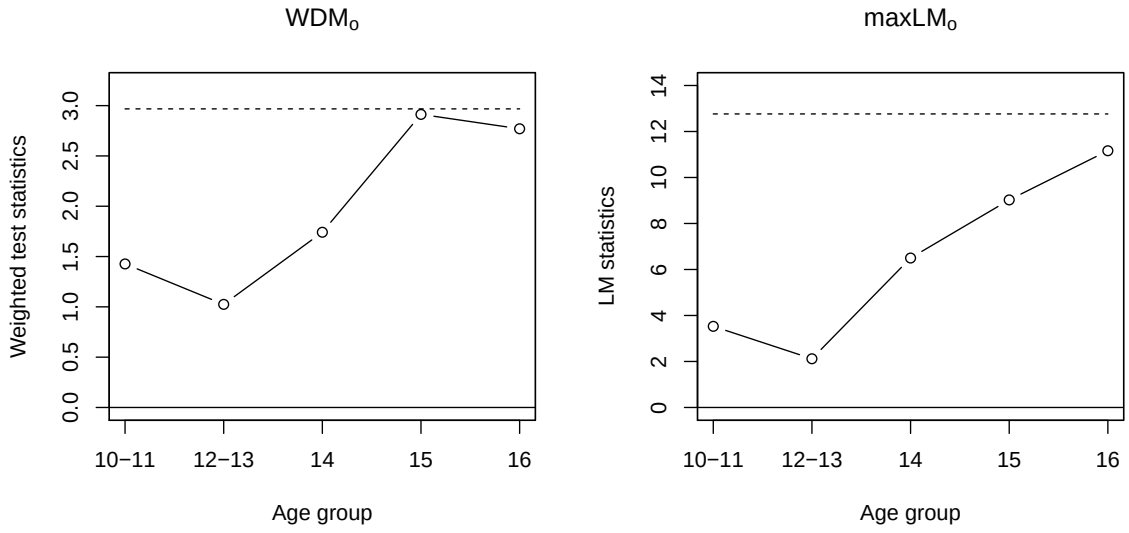

Figure 1: Fluctuation processes for the WDM_o statistic (left panel) and the max LM_o statistic (right panel).

`1:6`). In the same way, `nproc` can be employed to simulate higher-dimensional fluctuation processes suitable for testing more parameters. One can re-use `mLMo` in this manner for further tests of the youth gratitude data. Critical values must be resimulated for new data, however, because they depend on the proportion of individuals observed at each level of the ordinal variable (denoted t_ℓ for Equation 7).

In the above sections, we have illustrated the score-based tests' computation in R. We suspect that the ordinal tests will be most popular with users, because measurement invariance tests are typically carried out across categories (ordered or not), as opposed to continuous variables. Thus, in the sections below, we conduct novel simulations to study the ordinal statistics' expected behavior in practice. In particular, we wish to study (i) the extent to which the ordinal statistics attribute measurement invariance violations to the correct parameter(s), and (ii) the extent to which the tests are robust to model misspecification. These issues are especially important to examine because SEMs are typically complex, with many inter-related parameters that may exhibit measurement invariance. Previous applications of score-based tests have typically focused on regression-like models with only a small number of parameters that may exhibit instability (e.g., Zeileis and Hornik 2007). Thus, the simulations here provide general evidence about the extent to which the tests accurately capture instabilities in complex models.

4. Simulation 1

In Simulation 1, we examined the extent to which the proposed tests are sensitive to the specific model parameter that violates measurement invariance. If, say, a factor loading violates measurement invariance, it is plausible that this violation impacts other parameter estimates, including factor covariances or the unique variance associated with the manifest

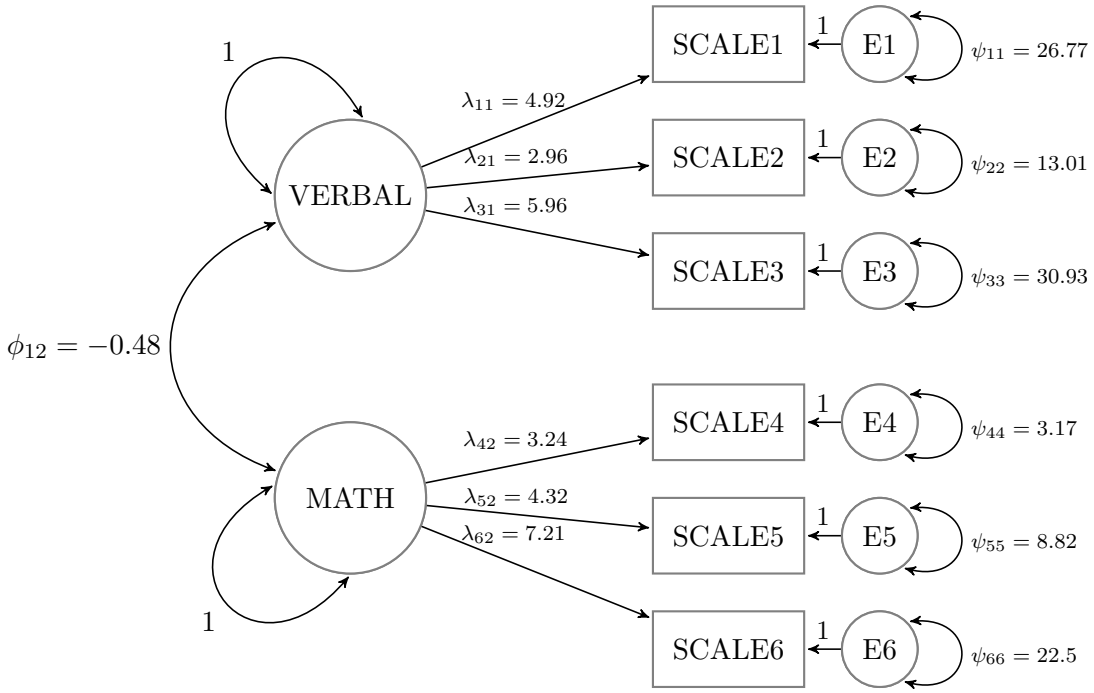

Figure 2: General model used for the simulations.

variable in question. Thus, the goal of the Simulation 1 is to examine the extent to which the proposed tests attribute the measurement invariance violation to the parameters that are truly in violation.

4.1. Methods

To examine these issues, we generated data from a two-factor model with three indicators each (see Figure 2). The measurement invariance violation occurred in one of three places: the factor loading associated with Scale 1 (λ_{11}), the unique variance associated with Scale 1 (ψ_{11}), or the factor covariance ϕ_{12}. We then tested for measurement invariance in five subsets of parameters: each of the three individual parameters noted above, all six factor loadings, and all six unique variances.

Power and type I error were examined across three sample sizes ($n = 120, 480, 960$), three numbers of categories ($m = 4, 8, 12$), and 17 magnitudes of invariance violations (described in the following). The measurement invariance violations began at level $1 + m/2$ of the auxiliary variable V and were consistent thereafter: individuals below level $1 + m/2$ of V deviated from individuals at or above level $1 + m/2$ by d times the parameters' asymptotic standard errors (scaled by $\sqrt{n}$), with $d = 0, 0.25, 0.5, ..., 4$. For each combination of sample size (n) $\times$ violation magnitude (d) $\times$ violating parameter $\times$ categories (m), 5000 datasets were generated and tested. Statistics from Equations 6, 7 and 8 were examined. As mentioned previously, Equation 8 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. In all conditions, we maintained equal sample sizes in each subgroup of the ordinal variable.

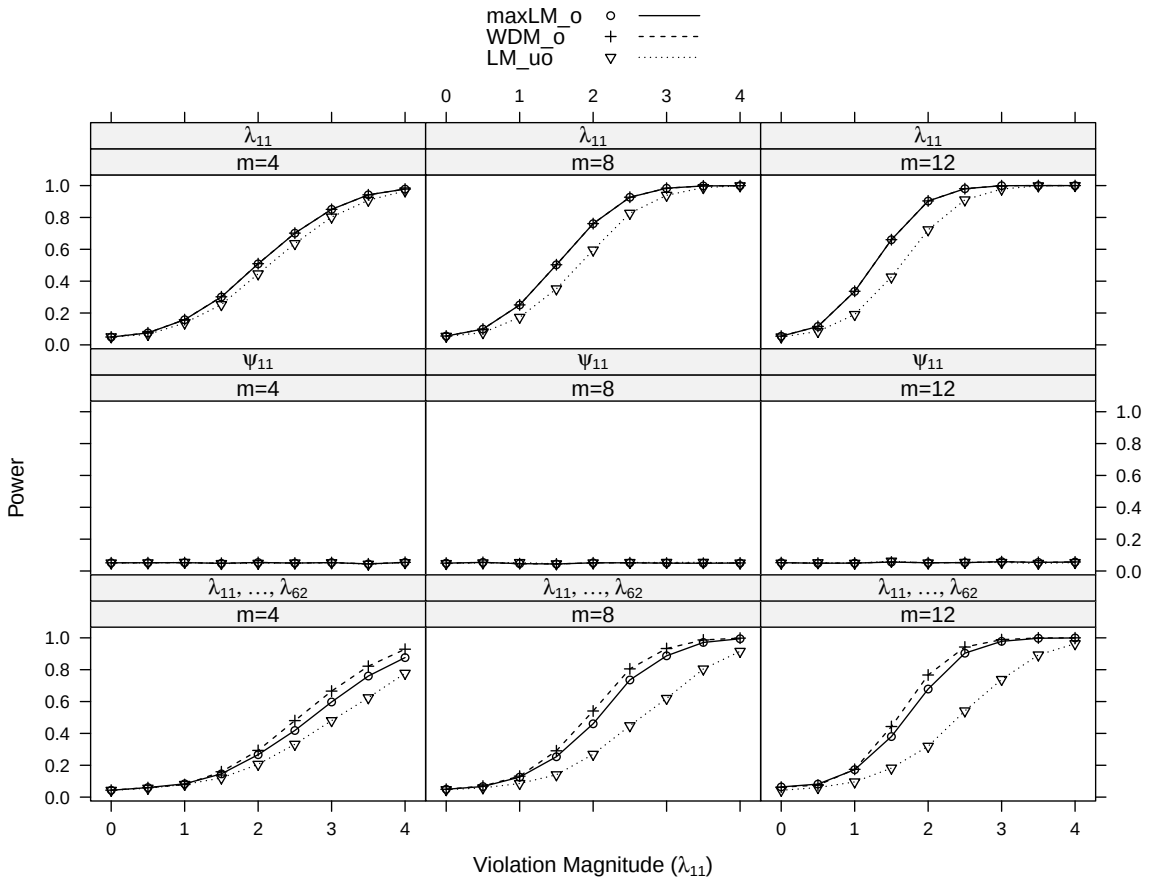

Figure 3: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 1. The parameter violating measurement invariance is λ_{11}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

4.2. Results

Full simulation results are presented in Figures 3 to 5. Figure 3 displays power curves as a function of violation magnitude in the factor loading λ_{11}, with the parameters being tested changing across rows, the number of levels m of the ordinal variable V across columns, and lines reflecting different test statistics. Figures 4 and Figure 5 display similar power curves when the factor covariance ϕ_{12} and error variance ϵ_{11} violate measurement invariance, respectively. In these figures, we generally show tests associated with parameters that exhibited nonzero power curves. For example, in Figure 3, the middle row shows that power for tests of ψ_{11} stays near zero for all values of m and d. Similar rows have been omitted from this figure and other figures.

Within each panel of Figure 3 to Figure 5, the three lines reflect the three test statistics. It is seen that the two ordinal statistics exhibit similar results, with $\max LM_{uo}$ 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

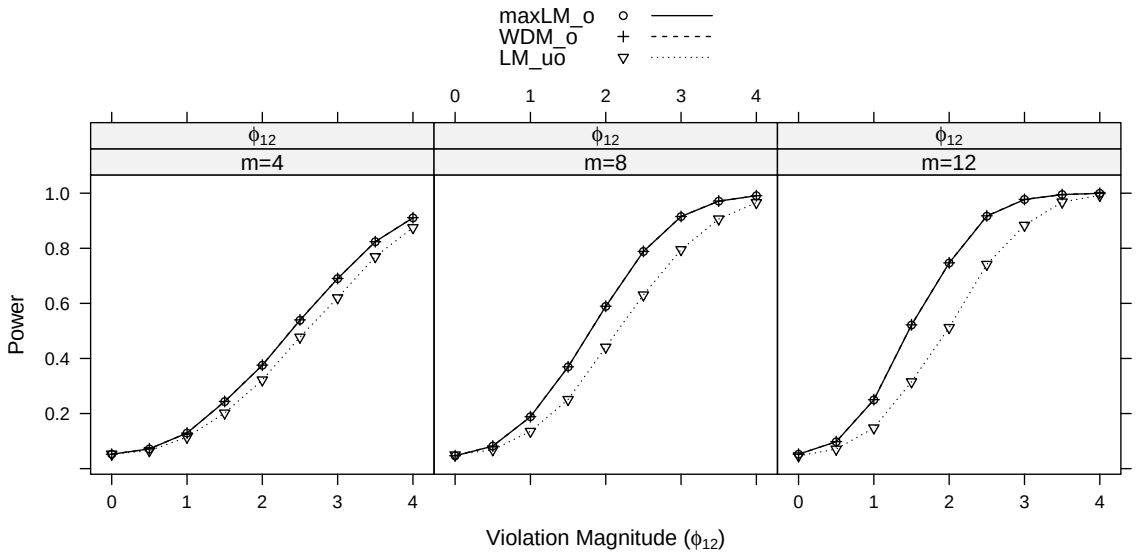

Figure 4: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m, and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 1. The parameter violating measurement invariance is ϕ_{12}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

is tested, WDM_o and $\max LM_o$ exhibit equivalent power curves. This is because, when only one parameter is tested, the statistics are equivalent.

From these figures, one generally observes that the tests isolate the parameter violating measurement invariance. Additionally, the tests have somewhat higher power to detect measurement invariance violations in the factor loading and factor covariance parameters, as opposed to the error variance parameter. Finally, simultaneous tests of all factor loadings or all error parameters result in decreased power, as compared to the situation where one tests only the violating parameter. This occurs because, in testing a subset of parameters (only one of which violates measurement invariance), we are effectively dampening the signal of a measurement invariance violation. This "dampening" effect is more apparent for the $\max LM_o$ statistic, because it involves a sum across all tested parameters (see Equation 7). Conversely, WDM_o takes the maximum over parameters (Equation 6), so that invariant parameters have no impact on this statistic.

In summary, we found that the proposed tests can attribute measurement invariance violations to the correct parameter. This provides evidence that, in practice, one can have confidence in the tests' abilities to locate the measurement invariance violation. Of course, this statement is qualified by the fact that, in this simulation, the model was correctly specified. In the following simulation, we examine the tests' performance in the likely situation of model misspecification.

5. Simulation 2

In Simulation 2, we examine the extent to which the results of Simulation 1 are robust to model misspecification. Specifically, we generate data from the factor analysis model used in

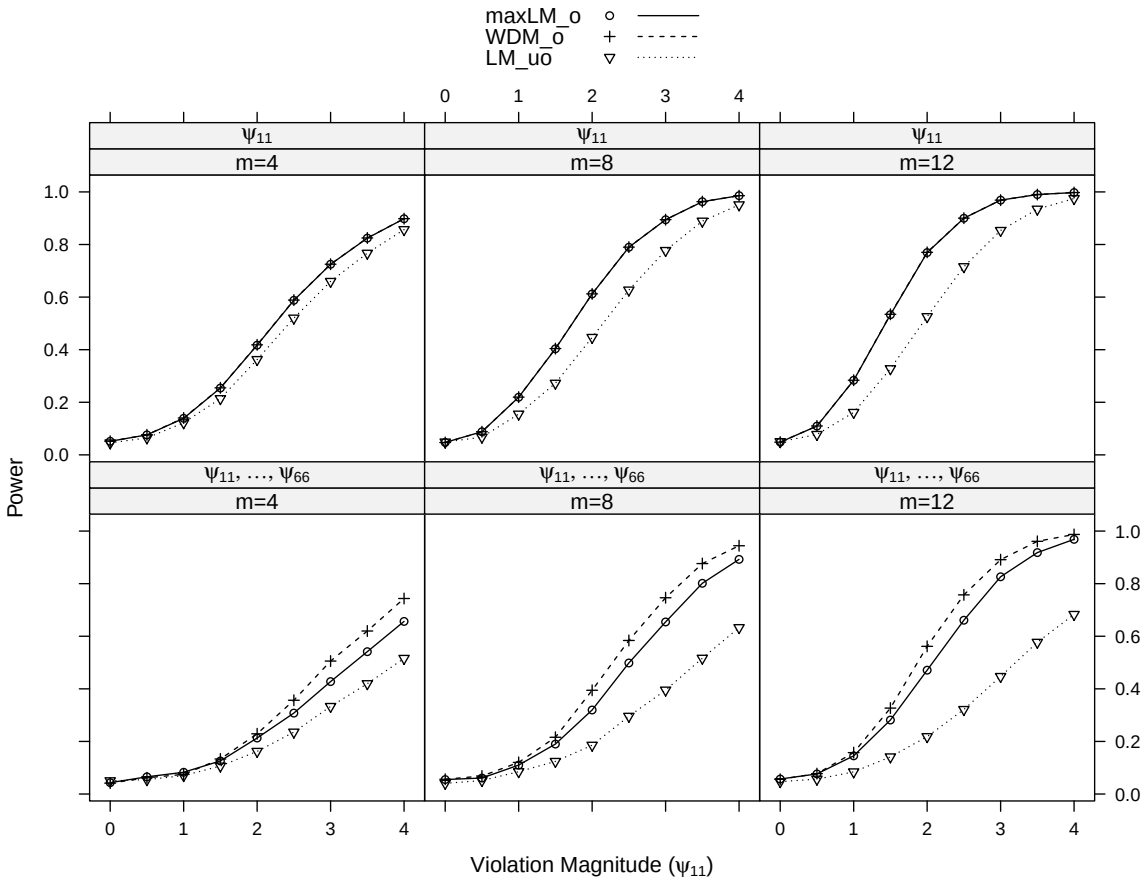

Figure 5: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 1. The parameter violating measurement invariance is ψ_{11}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

the previous section, except that the model contains an extra loading from the second factor to Scale 1. The estimated model matches that displayed in Figure 2, however, resulting in model misspecification. The goal of this simulation is to examine the proposed statistics' power to detect measurement invariance violations (and to attribute the violation to the correct parameter) under this misspecification.

5.1. Method

A measurement invariance violation could occur in each of the three parameters from Simulation 1 (factor loading, factor covariance, and unique variance), and a violation could also occur in the extra, unmodeled loading. In each condition, a single parameter exhibited the violation. Sample size and magnitude of measurement invariance violation were manipulated in the same way as they were in Simulation 1. The tested parameters were also the same as Simulation 1.

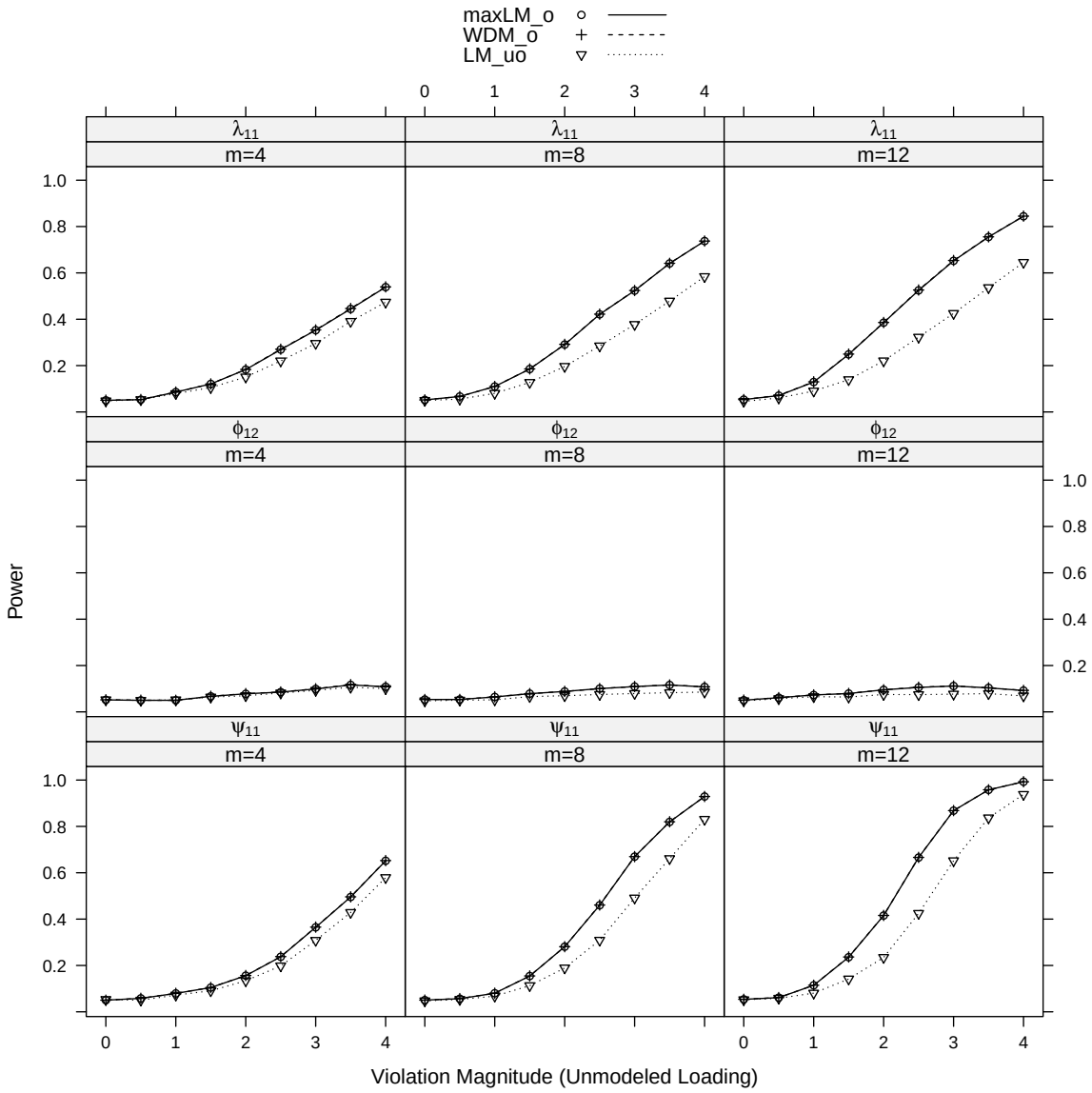

Figure 6: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 2. The parameter violating measurement invariance is the unmodeled loading. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

5.2. Results

Results of primary interest are conditions where the unmodeled loading violates measurement invariance. A subset of results is displayed in Figure 6. One can generally observe that tests of the first loading and unique variance exhibited high "power," which is actually a high type I error rate here. This type I error also demonstrated when testing all loadings and all unique variances, as shown in the appendix. Tests associated with the factor covariance did not

demonstrate this error, however. In terms of specific statistic performance, $\max LM_o$ and WDM_o demonstrated higher type I error than LM_{uo} in each panel, especially with increasing levels. This is likely because the unmodeled loading's non-invariance was monotonic with V; if it were not monotonic, we would expect LM_{uo} to have higher type I error.

When the parameter violating measurement invariance was modeled, results were generally the same as Simulation 1. When the modeled factor loading, λ_{11}, violated measurement invariance, the statistics were generally able to pick up the violation despite the misspecification. Similar results were observed when the unique variance and factor covariance parameters violated measurement invariance; these results are all shown in the Appendix. In particular, the power of ordered statistics $\max LM_o$ test and WDM_o test showed higher power than unordered LM_{uo} in each of these panels.

In summary, the proposed test statistics appear robust to unmodeled loading parameters, so long as the unmodeled loading does not violate measurement invariance. If the unmodeled loading does violate measurement invariance, the tests can still detect measurement invariance violations. The violations are assigned to modeled parameters that do not violate measurement invariance, however. The impacted parameters include the error variance and other loadings associated with the manifest variable that has an unmodeled loading. Thus, as for other tests of measurement invariance, it is important to study the extent to which the hypothesized model includes all parameters of importance. None of these tests (score-based or otherwise) inform the researcher of model misspecification.

6. General discussion

In this paper, we first described a novel family of test statistics for measurement invariance and illustrated their use via the R packages **lavaan** and **strucchange**. Next, we examined these statistics' abilities to identify the parameter violating measurement invariance under well-specified and misspecified models. We found that the proposed statistics could generally isolate the model parameter violating measurement invariance, so long as the violating parameter is included in the model.

In the remainder of the paper, we first compare these tests to traditional tests of measurement invariance. We then discuss test extension to other fit functions and to other specialized models.

6.1. Applications

Many of the applications in this volume, along with many measurement invariance applications in general, focus on testing across unordered categories such as nations or gender. As discussed earlier in this paper, the score-based tests for unordered categories are essentially equivalent to the usual likelihood ratio test. Given a measurement invariance violation across these unordered categories, however, researchers typically wish to know why the violation occurred. At this point, researchers may examine education level, socioeconomic status, income levels, and so on across the unordered categories. These variables are often ordinal or continuous in nature, so that the family of tests described in this paper are applicable. This is a first step towards describing why measurement invariance violations occur, as opposed to simply detecting measurement invariance violations. The tests described here are convenient for this purpose, as they do not require a new model to be estimated for each ordinal variable.

Instead, each ordinal variable defines an ordering of observations, which in turn yields a test statistic that is specific to that ordinal variable.

6.2. Extension

In this paper, we focused on testing for measurement invariance in factor analysis models that assume multivariate normality and that are estimated via maximum likelihood (ML). The family of tests described here generally apply to estimation methods that maximize/minimize a fit function, however (see Zeileis and Hornik 2007), so they are potentially applicable to alternative SEM discrepancy functions such as generalized least squares (e.g., Browne and Arminger 1995). Score calculation for these alternative discrepancy functions has not been implemented (to our knowledge), though the calculation could be implemented. Test statistic calculation and inference would then proceed in exactly the same manner as the calculation and inference illustrated in this paper.

In addition to alternative fit functions, the tests can be extended to other models estimated via ML. Of primary interest to the topic of measurement invariance, the tests can be extended to item response models to examine differential item functioning. In particular, Strobl, Kopf, and Zeileis (2013) studied application of these tests to the Rasch model, using them as the basis of a recursive partioning procedure that segments subgroups of individuals who exhibit DIF. Further study and extension of these tests for IRT seem warranted.

Computational details

All results were obtained using the R system for statistical computing (R Core Team 2013), version 3.0.1, employing the add-on package **lavaan** 0.5-14 (Rosseel 2012) for fitting of the factor analysis models and **strucchange** 1.5-0 (Zeileis *et al.* 2002; Zeileis 2006a) for evaluating the parameter instability tests. R and both packages are freely available under the General Public License 2 from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`. Replication R code for all results is available at `http://semtools.R-Forge.R-project.org/`.

Acknowledgments

This work was supported by National Science Foundation grant SES-1061334.

References

Bentler PM (1990). "Comparative Fit Indexes in Structural Models." *Psychological Bulletin*, **107**, 238–246.

Bentler PM, Bonett DG (1980). "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures." *Psychological Bulletin*, **88**, 588–606.

Browne MW, Arminger G (1995). "Specification and Estimation of Mean- and Covariance-Structure Models." In G Arminger, CC Clogg, ME Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pp. 185–249. Plenum Press, New York.

Froh JJ, Fan J, Emmons RA, Bono G, Huebner ES, Watkins P (2011). "Measuring Gratitude in Youth: Assessing the Psychometric Properties of Adult Gratitude Scales in Children and Adolescents." *Psychological Assessment*, **23**, 311–324.

Hansen BE (1997). "Approximate Asymptotic p Values for Structural-Change Tests." *Journal of Business & Economic Statistics*, **15**, 60–67.

Hothorn T, Zeileis A (2008). "Generalized Maximally Selected Statistics." *Biometrics*, **64**(4), 1263–1269.

Merkle EC, Fan J, Zeileis A (2013). "Testing for Measurement Invariance with Respect to an Ordinal Variable." *Psychometrika*. Forthcoming.

Merkle EC, Zeileis A (2013). "Tests of Measurement Invariance without Subgroups: A Generalization of Classical Methods." *Psychometrika*, **78**(1), 59–82.

Muthén B, Asparouhov T (2013). "BSEM Measurement Invariance Analysis." *Technical report*, Mplus Web Note 17.

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

Rosseel Y (2012). "**lavaan**: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36. URL http://www.jstatsoft.org/v48/i02/.

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

Strobl C, Kopf J, Zeileis A (2013). "A New Method for Detecting Differential Item Functioning in the Rasch Model." *Psychometrika*. Forthcoming.

van de Schoot R, Lugtig P, Hox J (2012). "A Checklist for Testing Measurement Invariance." *European Journal of Developmental Psychology*, **9**(4), 486–492.

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. URL 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**(2), 1–38. URL http://www.jstatsoft.org/v07/i02/.

Zeileis A, Strobl C, Wickelmaier F (2013). ***psychotools**: Infrastructure for Psychometric Modeling*. R package version 0.1-5.

A. Additional simulation results

This appendix contains additional results from Simulation 2, where there existed model misspecification (lacking one loading from Scale 1 to Math). Figure 7 to Figure 10 display results when the unmodeled loading, the loading λ_{11}, the covariance ϕ_{12}, and the error term ψ_{11} violate invariance, respectively. Results are only shown for simulation conditions exhibiting power curves that increased from zero. These results were generally the same as the Simulation 1 results.

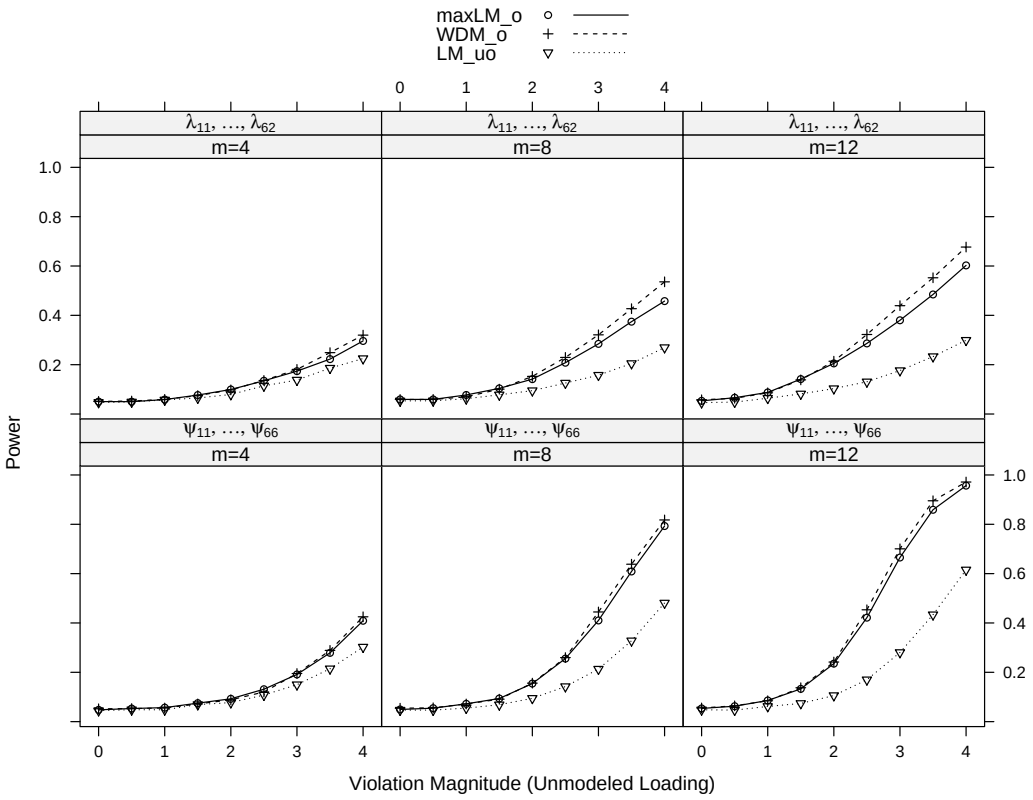

Figure 7: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 2. The parameter violating measurement invariance is the unmodeled loading. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

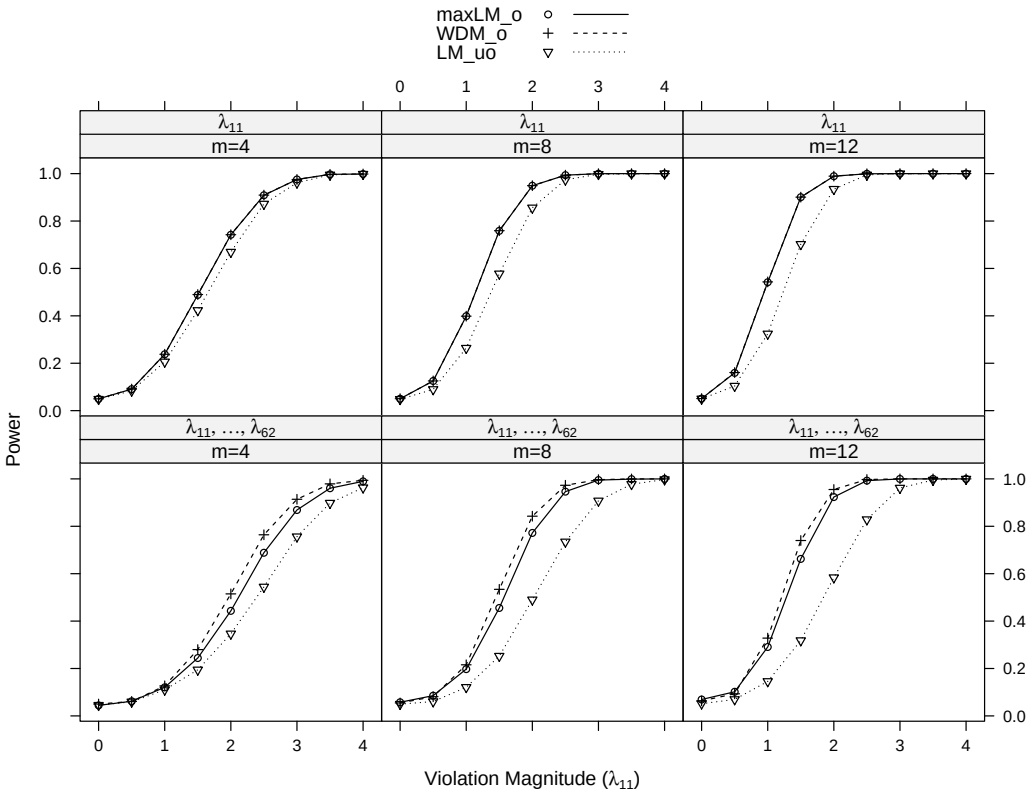

Figure 8: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 2. The parameter violating measurement invariance is λ_{11}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

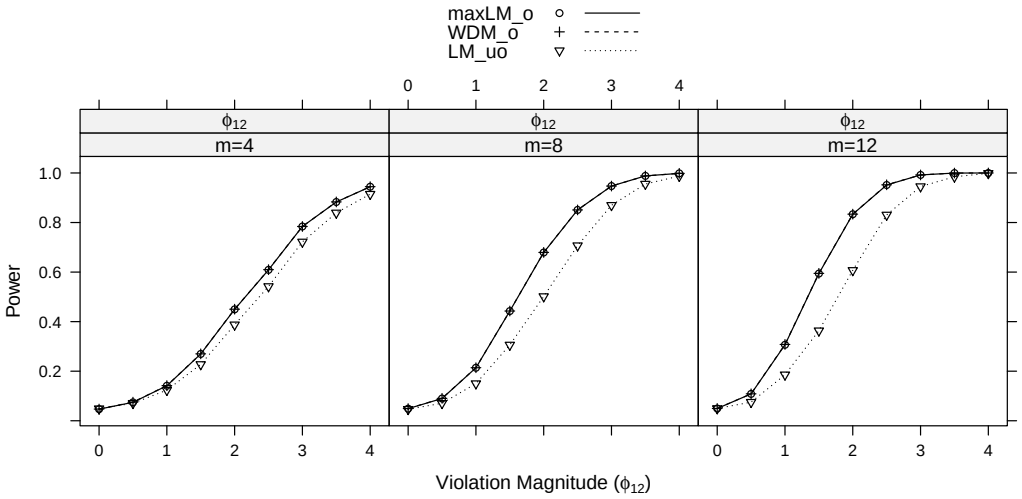

Figure 9: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 2. The parameter violating measurement invariance is ϕ_{12}. Panel labels denote the parameter being tested and the number of levels of the ordinal variable m.

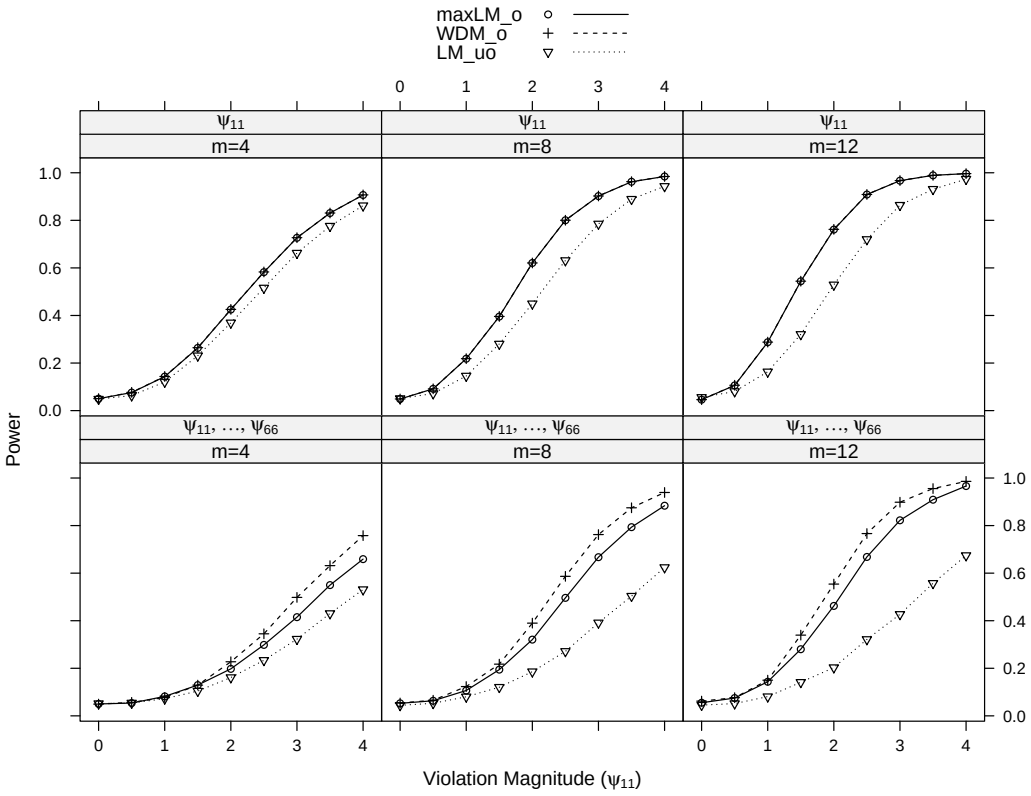

Figure 10: Simulated power curves for $\max LM_o$, WDM_o, and LM_{uo} across three levels of the ordinal variable m and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 2. The parameter violating measurement invariance is ψ_{11}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable m.

Affiliation:

Ting Wang, Edgar C. Merkle
Department of Psychological Sciences
University of Missouri
Columbia, MO 65211, United States of America
E-mail: twb8d@mail.missouri.edu, merklee@missouri.edu

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: http://eeecon.uibk.ac.at/~zeileis/

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

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

2013-33 **Ting Wang, Edgar C. Merkle, Achim Zeileis:** Score-based tests of measurement invariance: Use in practice

2013-32 **Jakob W. Messner, Georg J. Mayr, Daniel S. Wilks, Achim Zeileis:** Extending extended logistic regression for ensemble post-processing: Extended vs. separate vs. ordered vs. censored

2013-31 **Anita Gantner, Kristian Horn, Rudolf Kerschbamer:** Fair division in unanimity bargaining with subjective claims

2013-30 **Anita Gantner, Rudolf Kerschbamer:** Fairness and efficiency in a subjective claims problem

2013-29 **Tanja Hörtnagl, Rudolf Kerschbamer, Rudi Stracke, Uwe Sunde:** Heterogeneity in rent-seeking contests with multiple stages: Theory and experimental evidence

2013-28 **Dominik Erharter:** Promoting coordination in summary-statistic games

2013-27 **Dominik Erharter:** Screening experts' distributional preferences

2013-26 **Loukas Balafoutas, Rudolf Kerschbamer, Matthias Sutter:** Second-degree moral hazard in a real-world credence goods market

2013-25 **Rudolf Kerschbamer:** The geometry of distributional preferences and a non-parametric identification approach

2013-24 **Nadja Klein, Michel Denuit, Stefan Lang, Thomas Kneib:** Nonlife ratemaking and risk management with bayesian additive models for location, scale and shape

2013-23 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian structured additive distributional regression

2013-22 **David Plavcan, Georg J. Mayr, Achim Zeileis:** Automatic and probabilistic foehn diagnosis with a statistical mixture model

2013-21 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis, Daniel S. Wilks:** Extending extended logistic regression to effectively utilize the ensemble spread

2013-20 **Michael Greinecker, Konrad Podczeck:** Liapounoff's vector measure theorem in Banach spaces *forthcoming in Economic Theory Bulletin*

2013-19 **Florian Lindner:** Decision time and steps of reasoning in a competitive market entry game

2013-18 **Michael Greinecker, Konrad Podczeck:** Purification and independence

2013-17 **Loukas Balafoutas, Rudolf Kerschbamer, Martin Kocher, Matthias Sutter:** Revealed distributional preferences: Individuals vs. teams

2013-16 **Simone Gobien, Björn Vollan:** Playing with the social network: Social cohesion in resettled and non-resettled communities in Cambodia

2013-15 **Björn Vollan, Sebastian Prediger, Markus Frölich:** Co-managing common pool resources: Do formal rules have to be adapted to traditional ecological norms?

2013-14 **Björn Vollan, Yexin Zhou, Andreas Landmann, Biliang Hu, Carsten Herrmann-Pillath:** Cooperation under democracy and authoritarian norms

2013-13 **Florian Lindner, Matthias Sutter:** Level-k reasoning and time pressure in the 11-20 money request game *forthcoming in Economics Letters*

2013-12 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

2013-11 **Thomas Stöckl:** Price efficiency and trading behavior in limit order markets with competing insiders *forthcoming in Experimental Economics*

2013-10 **Sebastian Prediger, Björn Vollan, Benedikt Herrmann:** Resource scarcity, spite and cooperation

2013-09 **Andreas Exenberger, Simon Hartmann:** How does institutional change coincide with changes in the quality of life? An exemplary case study

2013-08 **E. Glenn Dutcher, Loukas Balafoutas, Florian Lindner, Dmitry Ryvkin, Matthias Sutter:** Strive to be first or avoid being last: An experiment on relative performance incentives.

2013-07 **Daniela Glätzle-Rützler, Matthias Sutter, Achim Zeileis:** No myopic loss aversion in adolescents? An experimental note

2013-06 **Conrad Kobel, Engelbert Theurl:** Hospital specialisation within a DRG-Framework: The Austrian case

2013-05 **Martin Halla, Mario Lackner, Johann Scharler:** Does the welfare state destroy the family? Evidence from OECD member countries

2013-04 **Thomas Stöckl, Jürgen Huber, Michael Kirchler, Florian Lindner:** Hot hand belief and gambler's fallacy in teams: Evidence from investment experiments

2013-03 **Wolfgang Luhan, Johann Scharler:** Monetary policy, inflation illusion and the Taylor principle: An experimental study

2013-02 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Tensions between the resource damage and the private benefits of appropriation in the commons

2013-01 **Jakob W. Messner, Achim Zeileis, Jochen Broecker, Georg J. Mayr:** Improved probabilistic wind power forecasts with an inverse power curve transformation and censored regression

2012-27 **Achim Zeileis, Nikolaus Umlauf, Friedrich Leisch:** Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond

2012-26 **Francisco Campos-Ortiz, Louis Putterman, T.K. Ahn, Loukas Balafoutas, Mongoljin Batsaikhan, Matthias Sutter:** Security of property as a public good: Institutions, socio-political environment and experimental behavior in five countries

2012-25 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Appropriation in the commons: variations in the opportunity costs of conservation

2012-24 **Edgar C. Merkle, Jinyan Fan, Achim Zeileis:** Testing for measurement invariance with respect to an ordinal variable *forthcoming in Psychometrika*

2012-23 **Lukas Schrott, Martin Gächter, Engelbert Theurl:** Regional development in advanced countries: A within-country application of the Human Development Index for Austria

2012-22 **Glenn Dutcher, Krista Jabs Saral:** Does team telecommuting affect productivity? An experiment

2012-21 **Thomas Windberger, Jesus Crespo Cuaresma, Janette Walde:** Dirty floating and monetary independence in Central and Eastern Europe - The role of structural breaks

2012-20 **Martin Wagner, Achim Zeileis:** Heterogeneity of regional growth in the European Union

2012-19 **Natalia Montinari, Antonio Nicolo, Regine Oexl:** Mediocrity and induced reciprocity

2012-18 **Esther Blanco, Javier Lozano:** Evolutionary success and failure of wildlife conservancy programs

2012-17 **Ronald Peeters, Marc Vorsatz, Markus Walzl:** Beliefs and truth-telling: A laboratory experiment

2012-16 **Alexander Sebald, Markus Walzl:** Optimal contracts based on subjective evaluations and reciprocity

2012-15 **Alexander Sebald, Markus Walzl:** Subjective performance evaluations and reciprocity in principal-agent relations

2012-14 **Elisabeth Christen:** Time zones matter: The impact of distance and time zones on services trade

2012-13 **Elisabeth Christen, Joseph Francois, Bernard Hoekman:** CGE modeling of market access in services

2012-12 **Loukas Balafoutas, Nikos Nikiforakis:** Norm enforcement in the city: A natural field experiment *forthcoming in European Economic Review*

2012-11 **Dominik Erharter:** Credence goods markets, distributional preferences and the role of institutions

2012-10 **Nikolaus Umlauf, Daniel Adler, Thomas Kneib, Stefan Lang, Achim Zeileis:** Structured additive regression models: An R interface to BayesX

2012-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** History repeating: Spain beats Germany in the EURO 2012 Final

2012-08 **Loukas Balafoutas, Glenn Dutcher, Florian Lindner, Dmitry Ryvkin:** The optimal allocation of prizes in tournaments of heterogeneous agents

2012-07 **Stefan Lang, Nikolaus Umlauf, Peter Wechselberger, Kenneth Harttgen, Thomas Kneib:** Multilevel structured additive regression

2012-06 **Elisabeth Waldmann, Thomas Kneib, Yu Ryan Yu, Stefan Lang:** Bayesian semiparametric additive quantile regression

2012-05 **Eric Mayer, Sebastian Rueth, Johann Scharler:** Government debt, inflation dynamics and the transmission of fiscal policy shocks *forthcoming in Economic Modelling*

2012-04 **Markus Leibrecht, Johann Scharler:** Government size and business cycle volatility; How important are credit constraints? *forthcoming in Economica*

2012-03 **Uwe Dulleck, David Johnston, Rudolf Kerschbamer, Matthias Sutter:** The good, the bad and the naive: Do fair prices signal good types or do they induce good behaviour?

2012-02 **Martin G. Kocher, Wolfgang J. Luhan, Matthias Sutter:** Testing a forgotten aspect of Akerlof's gift exchange hypothesis: Relational contracts with individual and uniform wages

2012-01 **Loukas Balafoutas, Florian Lindner, Matthias Sutter:** Sabotage in tournaments: Evidence from a natural experiment *published in Kyklos*

University of Innsbruck

Working Papers in Economics and Statistics

2013-33

Ting Wang, Edgar C. Merkle, Achim Zeileis

Score-based tests of measurement invariance: Use in practice

Abstract
In this paper, we consider a family of recently-proposed measurement invariance tests that are based on the scores of a fitted model. This family can be used to test for measurement invariance w.r.t. a continuous auxiliary variable, without pre-specification of subgroups. Moreover, the family can be used when one wishes to test for measurement invariance w.r.t. an ordinal auxiliary variable, yielding test statistics that are sensitive to violations that are monotonically related to the ordinal variable (and less sensitive to non-monotonic violations). The paper is specifically aimed at potential users of the tests who may wish to know (i) how the tests can be employed for their data, and (ii) whether the tests can accurately identify specific models parameters that violate measurement invariance (possibly in the presence of model misspecification). After providing an overview of the tests, we illustrate their general use via the R packages lavaan and strucchange. We then describe two novel simulations that provide evidence of the tests' practical abilities. As a whole, the paper provides researchers with the tools and knowledge needed to apply these tests to general measurement invariance scenarios.

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