Running head: TESTS OF MEASUREMENT INVARIANCE WITHOUT SUBGROUPS

Tests of measurement invariance without subgroups: A generalization of classical methods

Edgar C. Merkle
University of Missouri

Achim Zeileis
Universität Innsbruck

Abstract

The issue of measurement invariance commonly arises in factor-analytic contexts, with methods for assessment including likelihood ratio tests, Lagrange multiplier tests, and Wald tests. These tests all require advance definition of the number of groups, group membership, and offending model parameters. In this paper, we study tests of measurement invariance based on stochastic processes of casewise derivatives of the likelihood function. These tests can be viewed as generalizations of the Lagrange multiplier test, and they are especially useful for: (1) identifying subgroups of individuals that violate measurement invariance along a continuous auxiliary variable without prespecified thresholds, and (2) identifying specific parameters impacted by measurement invariance violations. The tests are presented and illustrated in detail, including an application to a study of stereotype threat and simulations examining the tests' abilities in controlled conditions.

Tests of measurement invariance without subgroups: A generalization of classical methods

The assumption that parameters are invariant across observations is a fundamental tenet of many statistical models. A specific type of parameter invariance, measurement invariance, has implications for the general design and use of psychometric scales. This concept is particularly important because violations can render the scales difficult to interpret. That is, if a set of scales violates measurement invariance, then individuals with the same "amount" of a latent variable may systematically receive different scale scores. This may lead researchers to conclude subgroup differences on a wide variety of interesting constructs when, in reality, the scales impact the magnitude of the differences. Further, it can be inappropriate to incorporate scales violating measurement invariance into structural equation models, where relationships between latent variables are hypothesized. Horn and McArdle (1992) concisely summarize the impact of these issues, stating "Lack of evidence of measurement invariance equivocates conclusions and casts doubt on theory in the behavioral sciences" (p. 141). Borsboom (2006) further notes that researchers often fail to assess whether measurement invariance holds.

In this paper, we apply a family of statistical tests based on stochastic processes to the assessment of measurement invariance. The tests are shown to have useful advantages over existing tests. We begin by developing a general framework for the tests, including discussion of theoretical results relevant to the proposed tests and comparison of the proposed tests to the existing tests. Next, we study the proposed tests' abilities through example and simulation. Finally, we discuss some interesting extensions of the tests. Throughout the manuscript, we use the term *test* to refer to a statistical test and the term *scale* to refer to a psychometric test or scale.

Framework

The methods proposed here are generally relevant to situations where the p-dimensional random variable X with associated observations $\boldsymbol{x}_i, i = 1, \dots, n$ is specified to arise from a model with density $f(\boldsymbol{x}_i; \boldsymbol{\theta})$ and associated joint log-likelihood

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

where $\boldsymbol{\theta}$ is some k-dimensional parameter vector that characterizes the distribution. The methods are applicable under very general conditions, essentially whenever standard assumptions for maximum likelihood inference hold (for more details see below). For the measurement invariance applications considered in this paper, we employ a factor analysis model with assumed multivariate normality:

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) \;=\; \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{1/2}} \; \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta})) \right\}, \tag{2}$$

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}_i) \;=\; -\frac{1}{2} \left\{ (\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta})) \;+\; \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| \;+\; p \log(2\pi) \right\}, \tag{3}$$

with model-implied mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. As pointed out above, the assumptions for the tests introduced here do not require this specific form of the likelihood, but it is presented for illustration due to its importance in practice.

Many expositions of factor analysis utilize the likelihood for the sample covariance matrix (instead of the likelihood of the individual observations $\boldsymbol{x}_i$), employing a Wishart distribution. However, in that approach the sample covariance matrix can also be disaggregated to the sum of its individual-specific contributions, leading essentially to a likelihood like (2) with multivariate normal observations $\boldsymbol{x}_i$. This situation is similar to that encountered in structural equation models with missing data (e.g., Wothke, 2000).

Within the general framework outlined above and under the usual regularity conditions (e.g., Ferguson, 1996), the model parameters $\boldsymbol{\theta}$ can be estimated by maximum likelihood (ML), i.e.,

$$\hat{\boldsymbol{\theta}} \;=\; \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}; x_1, \ldots, x_n), \tag{4}$$

or equivalently by solving the first order conditions

$$\sum_{i=1}^{n} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_i) \;=\; 0, \tag{5}$$

where

$$\boldsymbol{s}(\boldsymbol{\theta}; x_i) \;=\; \left(\frac{\partial \ell(\boldsymbol{\theta}; x_i)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}; x_i)}{\partial \theta_k} \right)^{\top}, \tag{6}$$

is the score function of the model, i.e., 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 $i = 1, \ldots, n$ essentially measures the extent to which each individual's likelihood is maximized.

One central assumption – sometimes made implicitly – is that the same model parameters $\boldsymbol{\theta}$ hold for all individuals $i = 1, \ldots, n$. If this is not satisfied, the estimates $\hat{\boldsymbol{\theta}}$ are typically not meaningful and cannot be easily interpreted. One potential source of deviation from this assumption is lack of measurement invariance, investigated in the following section.

Tests of Measurement Invariance

In general terms, a set of scales is defined to be measurement invariant with respect to an auxiliary variable V if:

$$f(x_i | t_i, v_i, \ldots) = f(\mathbf{x}_i | t_i, \ldots), \tag{7}$$

where x_i is the data vector for individual i, t_i is the latent variable for individual i that the scales purport to measure, and f is the model's distributional form (Mellenbergh, 1989). We adopt a parametric, factor-analytic framework here, so that the above equation being true for all V implies that the measurement parameters are equal across individuals and, thus, do not vary with any V (Meredith, 1993).

To frame this as a formal hypothesis, we assume that – in principle – Model (1) holds for all individuals but with a potentially individual-specific parameter vector $\boldsymbol{\theta}_i$. The null hypothesis of measurement invariance is then equivalent to the null hypothesis of parameter constancy

$$H_0: \ \boldsymbol{\theta}_i = \boldsymbol{\theta}_0, \quad (i = 1, \dots, n), \tag{8}$$

which should be tested against the alternative that the parameters are some nonconstant function $\boldsymbol{\theta}(\cdot)$ of the variable V with observations $v_1, \dots, v_n$, i.e.,

$$H_1: \ \boldsymbol{\theta}_i = \boldsymbol{\theta}(v_i), \quad (i = 1, \dots, n). \tag{9}$$

where the pattern $\boldsymbol{\theta}(V)$ of deviation from measurement invariance is typically not known (exactly) in practice. If it were (see below for some concrete examples), then standard inference methods – such as likelihood ratio, Wald, or Lagrange multiplier tests – could be employed. However, if the pattern is unknown, it is difficult to develop a single test that is well-suited for all conceivable patterns. But it is possible to derive a family of tests so that representatives from this family are well-suited for a wide range of possible patterns. One pattern of particular interest involves V dividing the individuals into two subgroups with different parameter vectors

$$H_1^*: \ \boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}^{(A)} & \text{if } v_i \le \nu, \\ \boldsymbol{\theta}^{(B)} & \text{if } v_i > \nu, \end{cases} \tag{10}$$

where $\boldsymbol{\theta}^{(A)} \neq \boldsymbol{\theta}^{(B)}$. This could pertain to two different age groups, income groups, genders, etc.

Note that even when adopting H_1^* as the relevant alternative, the pattern $\boldsymbol{\theta}(V)$ is not completely specified unless the cutpoint ν is known in advance (i.e., unless there is observed heterogeneity). In this situation, all individuals can be grouped based on V, and we can apply standard theory: nested multiple group models (e.g., Jöreskog, 1971; Bollen, 1989) coupled with likelihood ratio (LR) tests are most common, although the asymptotically equivalent Lagrange multiplier (LM) tests (also known as score tests) and Wald tests may also be used for this purpose (see Satorra, 1989). If ν is unknown (as is often the case for continuous V), however, then there is unobserved heterogeneity and standard theory is not easily applied. Nonstandard inference methods, such as those proposed in this paper, are then required.

In the following section, we describe the standard approaches to testing measurement invariance with ν known. We then contrast these approaches with the tests proposed in this paper. We assume throughout that the observations $i = 1, \dots, n$ are ordered with respect to the random variable V of interest such that $v_1 \le v_2 \le \dots \le v_n$. We also assume that the measurement model is correctly specified, as is implicitly assumed under traditional measurement invariance approaches. In particular, violations of normality may lead to spurious results in the proposed tests just as they do in other approaches (e.g., Bauer & Curran, 2004).

Likelihood Ratio, Wald, and Lagrange Multiplier Test for Fixed Subgroups

To employ the LR test for assessing measurement invariance, model parameters are estimated separately for a certain number of subgroups of the data (with some parameters potentially restricted to be equal across subgroups). For ease of exposition, we describe the case where there are no such parameter restrictions; as shown in the example and simulation below, however, it is straightforward to extend all methods to the more general case. After fitting the model to each subgroup, the sum of maximized likelihoods from the subgroups are compared with the original maximized full-sample likelihood in a χ^2 test. For the special case of two subgroups, the alternative H_1^* from (10) with fixed and prespecified ν is adopted and the null hypothesis H_0 from (8) reduces to $\boldsymbol{\theta}^{(A)} = \boldsymbol{\theta}^{(B)}$. The parameter estimates $\hat{\boldsymbol{\theta}}^{(A)}$ can then be obtained from the observations $i = 1, \ldots, m$, say, for which $v_i \leq \nu$. Analogously, $\hat{\boldsymbol{\theta}}^{(B)}$ is obtained by maximizing the likelihood for the observations $i = m+1, \ldots, n$, for which $v_i > \nu$. The LR test statistic for the given threshold ν is then

$$LR(\nu) \; = \; -2\left[\ell(\hat{\boldsymbol{\theta}}; x_1, \ldots, x_n) \; - \; \{\ell(\hat{\boldsymbol{\theta}}^{(A)}; x_1, \ldots, x_m) + \ell(\hat{\boldsymbol{\theta}}^{(B)}; x_{m+1}, \ldots, x_n)\}\right], \tag{11}$$

which, when the null hypothesis holds, has an asymptotic χ^2 with degrees of freedom equal to the number of parameters in $\boldsymbol{\theta}$.

Analogously to the LR test, the Wald test and LM test can be employed to test the null hypothesis H_1^* for a fixed threshold ν. For the Wald test, the idea is to compute the Wald statistic $W(\nu)$ as a quadratic form in $\hat{\boldsymbol{\theta}}^{(A)} - \hat{\boldsymbol{\theta}}^{(B)}$, utilizing its estimated covariance matrix for standardization. For the LM test, the LM statistic $LM(\nu)$ is a quadratic form in $\boldsymbol{s}(\hat{\boldsymbol{\theta}}; x_1, \ldots, x_m)$ and $\boldsymbol{s}(\hat{\boldsymbol{\theta}}; x_{m+1}, \ldots, x_n)$. Thus, the three tests all assess differences that should be zero under H_0: for the LR test the difference of maximized likelihoods; for the Wald test, the difference of parameter estimates; and for the LM test, the differences of likelihood scores from zero. In the LR case, the parameters have to be estimated under both the null hypothesis and alternative. Conversely, the Wald case requires only the estimates under the alternative, while the LM case requires only the estimates under the null hypothesis.

Extensions for Unknown Subgroups

For assessing measurement invariance in psychometric models, the major limitation of the three tests is that the potential subgroups have to be known in advance. Even if the scales are known to violate measurement invariance w.r.t. the variable V, the threshold ν from (10) is often unknown in practice. For example, if V represents yearly income, there are many possible values of ν that could be used to divide individuals into poorer and richer groups. The ultimate ν that we choose could potentially impact our conclusions about whether or not a scale is measurement invariant, in the same way that dichotomization of continuous variables impacts general psychometric analyses (MacCallum, Zhang, Preacher, & Rucker, 2002).

Instead of choosing a specific ν, a natural idea is to compute $LR(\nu)$ for each possible value in some interval $[\underline{\nu}, \overline{\nu}]$ and reject if their maximum

$$\max_{\nu \in [\underline{\nu}, \overline{\nu}]} LR(\nu) \tag{12}$$

becomes large. Note that this corresponds to maximizing the likelihood w.r.t. an additional parameter, namely ν. Hence, the asymptotic distribution of the maximum LR statistic is not χ^2 anymore. Andrews (1993) showed that the asymptotic distribution is in fact tractable but nonstandard. Specifically, the asymptotic distribution of (12) is the maximum of a certain tied-down Bessel process whose specifics also depend on the minimal and maximal thresholds $\underline{\nu}$ and $\overline{\nu}$, respectively. See Andrews for the original results and further references, and see below for more details on the results' application to measurement invariance.

Analogously, one can consider $\max W(\nu)$ and $\max LM(\nu)$, respectively, which both have the same asymptotic properties as $\max LR(\nu)$ and are asymptotically equivalent (Andrews, 1993). From a computational perspective, the $\max LM(\nu)$ test is particularly convenient because it requires just a single set of estimated parameters $\hat{\boldsymbol{\theta}}$ which is employed for all thresholds ν in $[\underline{\nu}, \overline{\nu}]$. The other two tests require reestimation of the subgroup models for each ν.

So far, the discussion focused on the alternative H_1^*: The maximum LR, Wald, and LM tests are designed for a situation where there is a single threshold at which all parameters in the vector $\boldsymbol{\theta}$ change. While this is plausible and intuitive in many applications, it would also be desirable to obtain tests that direct their power against other types of alternatives, i.e., against H_1 with other patterns $\boldsymbol{\theta}(V)$. For example, the parameters may fluctuate randomly or there might be multiple thresholds at which the parameters change. Alternatively, only one (or just a few) of the parameters in the vector $\boldsymbol{\theta}$ may change while the remaining parameters are constant (a common occurrence in psychometric models). To address such situations in a unified way, the next section contains a general framework for testing measurement invariance along a (continuous) variable V that includes the maximum LM test as a special case.

Stochastic Processes for Measurement Invariance

As discussed above, factor analysis models are typically estimated by fitting the model to all $i = 1, \dots, n$ individuals, assuming that the parameter vector $\boldsymbol{\theta}$ is constant across individuals. Having estimated the parameters $\hat{\boldsymbol{\theta}}$, the goal is to check that all subgroups of individuals conform with the model (for all of the parameters). Hence, some measure of model deviation or residual is required that captures the lack of fit for the i-th individual at the j-th parameter ($i = 1, \dots, n$, $j = 1, \dots, k$). A natural measure – that employs the ideas of the LM test – is $\boldsymbol{s}(\hat{\boldsymbol{\theta}}; x_i)_j$: the j-th component of the contribution of the i-th observation to the score function. By construction, the sum of the score contributions over all individuals is zero for each component; see (5). Moreover, if there are no systematic deviations, the score contributions should fluctuate randomly around zero. Conversely, the score contributions should be shifted away from zero for subgroups where the model does not fit.

Therefore, to employ this quantity for tests of measurement invariance against alternatives of type (9), we need to overcome two obstacles: (1) make use of the ordering of the observations w.r.t. V because we want to test for changes "along" V; (2) account for potential correlations between the k components of the parameters to be able to detect which parameter(s) change (if any).

Theory

The test problem of the null hypothesis (8) against the alternatives (9) and (10), respectively, has been studied extensively in the statistics and econometrics literature under the label "structural change tests" (see e.g., Brown, Durbin, & Evans, 1975; Andrews, 1993) where the focus of interest is the detection of parameter instabilities of time series models "along" time. Specifically, it has been shown (e.g., Nyblom, 1989; Hansen, 1992; Hjort & Koning, 2002; Zeileis & Hornik, 2007) that cumulative sums of the empirical scores follow specific stochastic processes, allowing us to use them to generally test measurement invariance. Here, we review some of the main results from that literature and adapt it to the specific challenges of factor analysis models. More detailed accounts of the underlying structural change methods include Hjort and Koning (2002) and Zeileis and Hornik (2007).

For application to measurement invariance, the most important theoretical result involves the fact that, under H_0 in (8), the *cumulative score process* converges to a specific asymptotic process. The k-dimensional cumulative score process is defined as

$$\boldsymbol{B}(t;\hat{\boldsymbol{\theta}}) \;=\; \hat{\boldsymbol{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; x_i) \qquad (0 \le t \le 1) \tag{13}$$

where $\lfloor nt \rfloor$ is the integer part of nt and $\hat{I}$ is some consistent estimate of the covariance matrix of the scores, e.g., their outer product or the observed information matrix. The k dimensions of the process arise from the fact that a separate cumulative score is maintained for each of the k model parameters. As the equation shows, the cumulative score process adds subsets of casewise score contributions across individuals along the ordering w.r.t. the variable V of interest. At $t = 1/n$, only the first individual's contribution enters into the summation; at $t = 2/n$, the first two individuals' contributions enter into the summation, etc., until $t = n/n$ where all contributions enter into the summation. Thus, due to (5), the cumulative score process always equals zero at $t = 0$ and returns to zero at $t = 1$. Furthermore, multiplication by $\hat{\boldsymbol{I}}^{-1/2}$ "decorrelates" the k cumulative score processes, such that each univariate process (i.e., each process for a single model parameter) is unrelated to (and asymptotically independent of) all other processes. Therefore, this cumulative process $\boldsymbol{B}(t;\hat{\boldsymbol{\theta}})$ accomplishes the challenges discussed at the beginning of this section: it makes use of the ordering of the observations by taking cumulative sums, and it decorrelates the contributions of the k different parameters.

Inference can then be based on an extension of the usual central limit theorem. Under the assumption of independence of individuals (implicit already in Equation 1) and under the usual ML regularity conditions (assuring asymptotic normality of $\hat{\boldsymbol{\theta}}$), Hjort and Koning (2002) show that

$$\boldsymbol{B}(\cdot;\hat{\boldsymbol{\theta}}) \;\overset{d}{\to}\; \boldsymbol{B}^0(\cdot), \tag{14}$$

where $\overset{d}{\to}$ denotes convergence in distribution and $\boldsymbol{B}^0(\cdot)$ is a k-dimensional Brownian bridge. In words, there are k cumulative score processes, one for each model parameter. This collection of processes follows a multidimensional Brownian bridge as score contributions accumulate in the summation from individual 1 (with lowest value of V) to individual n (with highest value of V).

The empirical cumulative score process from (13) can also be viewed as 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. That is, assuming that individuals are ordered by V, the first row of $\boldsymbol{B}(\hat{\boldsymbol{\theta}})$ corresponds to the first individual's decorrelated scores. The second row of $\boldsymbol{B}(\hat{\boldsymbol{\theta}})$ corresponds to the sum of the first two individuals' scores, etc., until the last row of $\boldsymbol{B}(\hat{\boldsymbol{\theta}})$ corresponds to the sum of all individuals' scores (which will be a row of zeroes). Under this setup, each column of $\boldsymbol{B}(\hat{\boldsymbol{\theta}})$ converges to a univariate Brownian bridge and pertains to a single factor analysis parameter. To carry out a test of H_0, the process/matrix needs to be aggregated to a scalar test statistic by collapsing across rows (individuals) and columns (parameters) of the matrix. The asymptotic distribution of this test statistic is then easily obtained by applying the same aggregation to the asymptotic Brownian bridge (Hjort & Koning, 2002; Zeileis & Hornik, 2007), so that corresponding p-values can be derived.

As argued above, no single aggregation function will have high power for all conceivable patterns of measurement invariance $\boldsymbol{\theta}(V)$, while any (reasonable) aggregation function will have non-trivial power under H_1. Thus, various aggregation strategies should be employed depending on which pattern $\boldsymbol{\theta}(V)$ is most plausible (because the exact pattern is typically unknown). A particularly agnostic aggregation strategy is to reject H_0 if any component of the cumulative score process $\boldsymbol{B}(t; \hat{\boldsymbol{\theta}})$ strays "too far" from zero at any time, i.e., if the double maximum statistic

$$DM = \max_{i=1,\dots,n} \max_{j=1,\dots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}|, \tag{15}$$

becomes large. In determining where in $\boldsymbol{B}(\hat{\boldsymbol{\theta}})$ this maximum occurred, we are able to determine threshold(s) of parameter change (over the individuals $i = 1, \dots, n$) and the parameter(s) affected by it (over $j = 1, \dots, k$). This test is especially useful for visualization, as the cumulative score process for each individual parameter can be displayed along with the appropriate critical value. For an example of this see Figure 3 (in the "Example" section) which shows the cumulative score processes for three factor loadings along with the critical values at 5% level.

However, taking maximums "wastes" power if many of the k parameters change at the same threshold, or if the score process takes large values for many of the n individuals (and not just a single threshold). In such cases, sums instead of maxima are more suitable for collapsing across parameters and/or individuals, because they combine the deviations instead of picking out only the single largest deviation. Thus, if the parameter instability $\boldsymbol{\theta}(V)$ affects many parameters and leads to many subgroups, sums of (absolute or squared) values should be used for collapsing both across parameters and individuals. On the other hand, if there is just a single threshold that affects multiple parameters, then the natural aggregation is by sums over parameters and then by the maximum over individuals. More precisely, the former idea leads to a Cramér-von Mises type statistic and the latter to the maximum LM statistic from the previous section:

$$\begin{aligned} CvM &= n^{-1} \sum_{i=1,\dots,n} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, & (16) \\ \max LM &= \max_{i=\underline{i},\dots,\overline{\imath}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, & (17) \end{aligned}$$

where the $\max LM$ statistic is additionally scaled by the asymptotic variance $t(1-t)$ of the process $\boldsymbol{B}(t, \hat{\boldsymbol{\theta}})$. It is equivalent to the $\max_\nu LM(\nu)$ statistic from the previous section, provided that the boundaries for the subgroups sizes $\underline{\imath}/\underline{\nu}$ and $\overline{\imath}/\overline{\nu}$ are chosen analogously. Further aggregation functions have been suggested in the structural change literature (see e.g., Zeileis, 2005; Zeileis, Shah, & Patnaik, 2010) but the three tests above are most likely to be useful in psychometric settings.

Finally, all tests can be easily modified to address the situation of so-called "partial structural changes" (Andrews, 1993). This refers to the case of some parameters being known to be stable, i.e., restricted to be constant across potential subgroups. Tests for potential changes/instabilities only in the k^* remaining parameters (from overall k parameters) are easily obtained by omitting those $k - k^*$ columns from $\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}$ that are restricted/stable, retaining only those k^* columns that are potentially unstable. This may be of special interest to those wishing to test specific types of measurement invariance, where subsets of model parameters are assumed to be stable.

Critical Values & p-values

As pointed out above, specification of the asymptotic distribution under H_0 for the test statistics from the previous section is straightforward: it is simply the aggregation of the asymptotic process $\boldsymbol{B}^0(t)$ (Hjort & Koning, 2002; Zeileis & Hornik, 2007). Thus, DM from (15) converges to $\sup_t ||\boldsymbol{B}^0(t)||_\infty$, where $|| \cdot ||_\infty$ denotes the maximum norm. Similarly, CvM from (16) converges to $\int_0^1 ||\boldsymbol{B}^0(t)||_2^2 dt$, where $|| \cdot ||_2$ denotes the Euclidean norm. Finally, $\max LM$ from (17) – and analogously the maximum Wald and LR tests – converges to $\sup_t (t(1-t))^{-1} ||\boldsymbol{B}^0(t)||_2^2$ (which can also be interpreted as the maximum of a tied-down Bessel process, as pointed out previously).

While it is easy to formulate these asymptotic distributions theoretically, it is not always easy to find closed-form solutions for computing critical values and p-values from them. In some cases – in particular for the double maximum test – such a closed-form solution is available from analytic results for Gaussian processes (see e.g., Shorack & Wellner, 1986). For all other cases, tables of critical values can be obtained from direct simulation (Zeileis, 2006) or in combination with more refined techniques such as response surface regression (Hansen, 1997).

The analytic solution for the asymptotic p-value of a DM statistic d is

$$P(DM > d \mid H_0) \overset{asy}{=} 1 - \left\{ 1 + 2 \sum_{h=1}^{\infty} (-1)^h \exp(-2h^2 d^2) \right\}^k . \qquad (18)$$

This combines the crossing probability of a univariate Brownian bridge (see e.g., Shorack & Wellner, 1986; Ploberger & Krämer, 1992) with a straightforward Bonferroni correction to obtain the k-dimensional case. The terms in the summation quickly go to zero as h goes to infinity, so that only some large finite number of terms (say, 100) need to be evaluated in practice.

For the Cramér-von Mises test statistic CvM, Nyblom (1989) and Hansen (1992) provide small tables of critical values which have been extended in the software provided by Zeileis (2006). Critical values for the distribution of the maximum LR/Wald/LM tests are provided by Hansen (1997). Note that the distribution depends on the minimal and maximal thresholds employed in the test.

Local Alternatives

Using results from Hjort and Koning (2002) and Zeileis and Hornik (2007), we can also capture the behavior of the process $\boldsymbol{B}(t, \hat{\boldsymbol{\theta}})$ under specific alternatives of parameter instability. In particular, we can assume that the pattern of deviation $\boldsymbol{\theta}(v_i)$ can be described as a constant parameter plus some non-constant deviation $\boldsymbol{g}(i/n)$:

$$\boldsymbol{\theta}(v_i) = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{g}(i/n). \tag{19}$$

In this case, the scores $\boldsymbol{s}(\boldsymbol{\theta}_0; x_i)$ from Equation (6) do not have zero expectation but rather

$$E[\boldsymbol{s}(\boldsymbol{\theta}_0; x_i)] = \boldsymbol{0} + n^{-1/2}\boldsymbol{C}(\boldsymbol{\theta}_0)\boldsymbol{g}(i/n). \tag{20}$$

The covariance matrix $\boldsymbol{C}(\boldsymbol{\theta})$ from the expected outer product of gradients is then

$$\boldsymbol{C}(\boldsymbol{\theta}_0) = E[\boldsymbol{s}(\boldsymbol{\theta}_0; x_i)\boldsymbol{s}(\boldsymbol{\theta}; x_i)'] \tag{21}$$

which, at $\boldsymbol{\theta}_0$, coincides with the expected information matrix.

Under the local alternative described above, Hjort and Koning (2002) show that the process $\boldsymbol{B}(t, \hat{\boldsymbol{\theta}})$ behaves asymptotically like

$$\boldsymbol{B}^0(t) + \hat{\boldsymbol{I}}^{-1/2}\boldsymbol{C}(\hat{\boldsymbol{\theta}})\boldsymbol{G}^0(t), \tag{22}$$

i.e., a zero-mean Brownian bridge plus a term with non-zero mean driven by $\boldsymbol{G}^0(t) = \boldsymbol{G}(t) - t\boldsymbol{G}(1)$, where $\boldsymbol{G}(t) = \int_0^t \boldsymbol{g}(y)dy$. Hence, unless the local alternative has $\boldsymbol{g}(t) \equiv \boldsymbol{0}$, the empirical process $\boldsymbol{B}(t, \hat{\boldsymbol{\theta}})$ will have a non-zero mean. Hence, the corresponding tests will have non-trivial power (asymptotically). We use these results in the “Simulation” section to describe the expected behavior of the Brownian bridge.

Locating the Invariance

If the employed parameter instability test detects a measurement invariance violation, the researcher is typically interested in identification of the parameter(s) affected by it and/or the associated threshold(s). As argued above, the double maximum test is particularly appealing for this because the k-dimensional empirical cumulative score process can be graphed along with boundaries for the associated critical values (see Figure 3). Boundary crossing then implies a violation of measurement invariance, and the location of the most extreme deviation(s) in the process convey threshold(s) in the underlying ordering V.

For the maximum LR/Wald/LM tests, it is natural to graph the sequence of LR/Wald/LM statistics along V, with a boundary corresponding to the critical value (see Figure 2 and the third row of Figure 4). Again, a boundary crossing signals a significant violation, and the peak(s) in the sequence of statistics conveys threshold(s). Note that, due to summing over all parameters, no specific parameter can be identified that is responsible for the violation. Similarly, neither component(s) nor threshold(s) can be formally identified for the Cramér-von Mises test. However, graphing of (transformations of) the cumulative score process may still be valuable for gaining some insights (see, e.g., the second row of Figure 4).

If a measurement invariance violation is detected by any of the tests, one may want to incorporate it into the model to account for it. The procedure for doing this typically depends on the type of violation $\boldsymbol{\theta}(V)$, and the visualizations discussed above often prove helpful in determining a suitable parameterization. In particular, one approach that is often employed in practice involves adoption of a model with one (or more) threshold(s) in all parameters (i.e., (10) for the single threshold case). In the multiple threshold case, their location can be determined by maximizing the corresponding segmented log-likelihood over all possible combinations of thresholds (Zeileis et al., 2010, adapt a dynamic programming algorithm to this task). For the single threshold case, this reduces to maximizing the segmented log-likelihood

$$\ell(\hat{\boldsymbol{\theta}}^{(A)}; x_1, \ldots, x_m) \; + \; \ell(\hat{\boldsymbol{\theta}}^{(B)}; x_{m+1}, \ldots, x_n) \tag{23}$$

over all values of m corresponding to possible thresholds ν (such that $v_m \leq \nu$ and $v_{m+1} > \nu$). As pointed out previously, this is equivalent to maximizing the LR statistic from (11) (with some minimal subgroup size typically imposed).

Formally speaking, the maximization of (23) – or equivalently (11) – yields an estimate $\hat{\nu}$ of the threshold in H_1^*. If H_1^* is in fact the true model, the peaks in the Wald/LM sequences and the cumulative score process, respectively, will occur at the same threshold asymptotically. However, in empirical samples, their location may differ (although often not by much).

These attributes give the proposed tests important advantages over existing tests, as existing measurement invariance methods cannot: (1) test measurement invariance for unknown ν, or (2) isolate specific parameters violating measurement invariance. In particular, Millsap (2005) cites “locating the invariance violation” as a major outstanding problem in the field. In the following sections, we demonstrate the proposed tests’ uses and properties via example and simulation. We first describe an example and simulation with artificial data, and we then provide an illustrative example with real data. Finally, we conclude the paper with a discussion of test extensions and general summary.

Example with Artificial Data

Consider a hypothetical battery of six scales administered to students aged 13 to 18 years. Three of the scales are intended to measure verbal ability, and three of the scales are intended to measure mathematical ability. We may observe a maturation effect in the resulting data, whereby the factor loadings for older students are larger than those for younger students. The researcher’s goal is to study whether the scales are measurement invariant with respect to age, which is taken to be the auxiliary variable V.

Method

To formally represent these ideas, we specify that the data arise from a factor analysis model with two factors. The base model, displayed in Figure 1, specifies that measurement invariance holds, with three scales arising from the verbal factor and three scales arising from the mathematical factor. For the measurement invariance violation, we specify that V (student age) impacts the values of verbal factor loadings in the model: if students are 16 through 18 years of age, then the factor loadings corresponding to the first factor $(\lambda_{11}, \lambda_{21}, \lambda_{31})$ reflect those in Figure 1. If students are 13 through 15 years of age, however, then the factor loadings corresponding to the first factor are three standard errors (= asymptotic standard errors divided by $\sqrt{n}$) lower than those in Figure 1. This violation states that the verbal ability scales lack measurement invariance with respect to age. For simplicity, we assume that the mathematical scales are invariant.

A sample of size 200 was generated from the model described above, and a test was conducted to examine measurement invariance of the three verbal scales. To carry out the test, a confirmatory factor analysis model (with the paths displayed in Figure 1) was fit to the data. Casewise derivatives and the observed information matrix were then obtained, and they were used to calculate the cumulative score process via (13). Finally, we obtained various test statistics and p-values from the cumulative score process. These include the double-max statistic from (15), the Cramér-von Mises statistic from (16), and the $\max LM$ statistic from (17).

As mentioned in the theory section, the tests give us the flexibility to study hypotheses of partial change. That is, we have the ability to test various subsets of parameters. For example, if we suspected that the verbal factor loadings lacked measurement invariance, we could test

$$H_0: \ (\lambda_{i,11} \ \lambda_{i,21} \ \lambda_{i,31}) = (\lambda_{0,11} \ \lambda_{0,21} \ \lambda_{0,31}), \ i = 1, \ldots, n, \tag{24}$$

where $(\lambda_{i,11} \ \lambda_{i,21} \ \lambda_{i,31})$ represent the verbal factor loading parameters for student i. Thus, here only $k^* = 3$ from the overall $k = 19$ model parameters are assessed (where the 19 parameters include six factor loadings, six unique variances, six intercepts, and one factor correlation). Alternatively, we can consider all $k^* = k = 19$ parameters, leading to a test of (8). We consider both of these tests below.

Results

Here, we first describe overall results and subsequently the identification of ν and isolation of model parameters violating measurement invariance.

Overall Results. Test statistics for the hypotheses (24) and (8) are displayed in Figure 4. Each panel displays a test statistic's fluctuation across values of student age, with the first column containing tests of (24) and the second column containing tests of (8). The solid horizontal lines represent critical values for $\alpha = 0.05$, and the Cramér-von Mises panels also contain a dashed line depicting the value of the test statistic CvM (test statistics for the others are simply the maxima of the processes). In other words, for panels in the first and third rows, (24) is rejected if the process crosses the horizontal line. For panels in the second row, (8) is rejected if the dashed horizontal line is higher than the solid horizontal line.

The figures convey information about several properties of the tests. First, all three tests are more powerful (and in this example significant) if we test only those parameters that are subject to instabilities. Conversely, if all 19 model parameters are assessed (including those that are in fact invariant), the power is decreased. This decrease, however, is less pronounced for the double-max test as it is more sensitive to fluctuations among a small subset of parameters (3 out of 19 here).

Figure 2 compares the max LM statistic (solid line) to the max LR statistic (dashed line) from (12), as applied to testing (8). (Note also that the visualization of the max LM in Figure 2 is identical to the bottom right panel of Figure 4.) The critical values for these two tests are identical, hence the single horizontal line. The figure shows that the two statistics are very similar to one another, with both maxima at the dotted vertical line. This is generally to be expected, because the two tests are asymptotically equivalent. The max LR statistic cannot be obtained from the empirical fluctuation process, however, so the factor analysis model must be refitted before and after each of the possible threshold values ν (i.e., here 320 model fits for 160 thresholds).

Interpretation of Test Results. As described above, the tests of (24) imply that the verbal scales lack measurement invariance. We can also use the tests to: (1) locate the threshold ν, and (2) isolate specific parameters that violate measurement invariance. For example, as described previously, information about the location of ν can be obtained by examining the peaks in Figure 4. For all six panels in the figure, the peaks occur near an age of 16.1. This agrees well with the true threshold of 16.0.

We previously noted that the double-max test is advantageous because it yields information about individual parameters violating measurement invariance. That is, it allows us to examine whether or not individual parameters' cumulative score processes lead one to reject the hypothesis of measurement invariance. Figure 3 shows the individual cumulative score processes for the three verbal factor loadings and the top left panel of Figure 4 shows the same processes aggregated over the three parameters. In both graphics, horizontal lines reflect the same critical value at $\alpha = 0.05$. The figure shows that the third parameter (i.e., λ_{31}) crosses the dashed line, so we would conclude a measurement invariance violation with respect to age for the third verbal test. The fact that the cumulative score process for the first and second loadings did not achieve the critical value represents a type II error, which implicitly brings into question the tests' power. We generally address the issue of power in the simulations below.

Simulation

In this section, we conduct a simulation designed to examine the tests' power and type I error rates in situations where measurement invariance violations are known to exist. Specifically, we generated data from the factor analysis model used in the previous section, with measurement invariance violations in the "verbal" factor loadings with respect to a continuous auxiliary variable.

We examine the power and error rates of the three tests used in the previous section: the double-max test, the Cramér-von Mises test, and the max LM test. We also compare tests involving only the $k^* = 3$ factor loadings lacking invariance with tests of all $k^* = k = 19$ model parameters. It is likely that power is higher when testing only the affected model parameters, but the affected parameters are usually unknown in practice. Thus, the $k^* = 3$ and $k^* = 19$ conditions reflect boundary cases, with the performance of tests involving other subsets of model parameters (e.g., all factor loadings) falling in between the boundaries. Finally, we also examine the tests' power across various magnitudes of measurement invariance violations.

Method

Data were generated from the same model and with the same type of measurement invariance violation as was described in the "Artificial Example" section. Sample size and magnitude of measurement invariance violation were manipulated to examine power: we examined power to detect invariance violations across four sample sizes ($n = 50, 100, 200, 500$) and 17 magnitudes of violations. These violations involved the younger students' values of $\{\lambda_{11}, \lambda_{21}, \lambda_{31}\}$ deviating from the older students' values by d times the parameters' asymptotic standard errors (scaled by $\sqrt{n}$), with $d = 0, 0.25, 0.5, \ldots, 4$. The $d = 0$ standard error condition was used to study type I error rate.

For each combination of sample size (n) × violation magnitude (d) × number of parameters being tested (k^*), 5,000 datasets were generated and tested. In each dataset, half the individuals had "low V" (e.g., 13–15 years of age) and half had "high V" (e.g., 16–18 years of age).

Using results discussed previously in the "Local Alternatives" section, we can derive the expected behavior of the univariate Brownian bridges for the parameters $\{\lambda_{11}, \lambda_{21}, \lambda_{31}\}$. For these parameters, we use a simple step function $g(t) = \mathrm{I}(t > 0.5)$ multiplied by the violation magnitude. Thus, $G(t) = \mathrm{I}(t > 0.5) \times (t - 0.5)$ and $G^0(t) = \{\mathrm{I}(t > 0.5) - 0.5\} \times t - \mathrm{I}(t > 0.5) \times 0.5$, both again multiplied by the violation magnitude. This implies that the mean under the alternative is driven by a function that is triangle-shaped, with a peak at the changepoint $t = 0.5$ for the parameters affected by the change (and equal to zero for all remaining parameters).

Results

Full simulation results are presented in Figure 5, and the underlying numeric values for a subset of the results are additionally displayed in Table 1. In describing the results, we largely refer to the figure.

Figure 5 displays power curves as a function of violation magnitude, with panels for each combination of sample size (n) × number of parameters being tested (k^*). Separate curves are drawn for the double-max test (solid lines), the Cramér-von Mises test (dashed lines), and the max LM test (dotted lines). One can generally observe that simultaneous tests of all 19 parameters result in decreased power, with the tests performing more similarly at the larger sample sizes. The tests distinguish themselves from one another when only the three factor loadings are tested, with the Cramér-von Mises test having the most power, followed by max LM, followed by the double-max test. This advantage decreases at larger sample sizes, and we also surmise that it decreases as extra parameters satisfying measurement invariance are included in the tests. At $n = 50$, generally regarded as a small sample size for factor analysis, the power functions break down and are not monotonic with respect to violation magnitude.

Table 1 presents a subset of the results displayed in Figure 5, but it is easier to see exact power magnitudes in the table. The table shows that the power advantage of the Cramér-von Mises test can be as large as 0.1, most notably when three parameters are being tested. It also shows that the Cramér-von Mises test generally is very close to its nominal type I error rate, with the double-max test being somewhat conservative and the max LM test being slightly liberal.

In summary, we found that the proposed tests have adequate power to detect measurement invariance violations in applied contexts. The Cramér-von Mises statistic exhibited the best performance for the data generated here, though more simulations are warranted to examine the generality of this finding in other models or other parameter constellations. In the discussion, we describe extensions of the tests in factor analysis and beyond.

Application: Stereotype Threat

Wicherts, Dolan, and Hessen (2005), henceforth WDH, utilized confirmatory factor models to study measurement invariance in a series of general intelligence scales. They were interested in the notion of *stereotype threat*, whereby stereotypes concerning a subgroup's ability adversely impact the subgroup's performance on tests of that ability. The authors specifically focused on stereotypes concerning ethnic minorities' performance on tests of general intelligence. In this section, we use the proposed tests to supplement the authors' original analyses.

Background

To study stereotype threat in a measurement invariance framework, Study 1 of Wicherts et al. (2005) involved 295 high school students completing three intelligence tests in two between-subjects conditions. Conditions were defined by whether or not students received primes about ethnic stereotypes prior to testing. To study the data, WDH employed a four-group, one-factor model with the three intelligence tests as manifest variables. The groups were defined by ethnicity (majority/minority) and by the experimental manipulation (received/did not receive stereotype prime). Results indicated that the intelligence tests lacked measurement invariance, with the minority group receiving stereotype primes being particularly different from the other three groups on the most difficult intelligence test. In the example below, we employ one of the models used by WDH. Our V is participants' aptitudes, as measured by their grade point averages (GPAs) which were unused in the original analyses.

Method

WDH tested a series of confirmatory factor models for various types of measurement invariance. We focus on the model used in their Step 5b (see pp. 703–704 of their paper), which involved across-group restrictions on the factor loadings, unique variances, intercepts, and factor variances. These parameters were typically restricted to be equal across groups, though a subset of the parameters were allowed to be group-specific upon examination of modification indices. The model provided a reasonable fit to the data, as judged by examination of χ^2, RMSEA (root-mean-square error of approximation), and CFI (comparative fit index) statistics. While the model included four groups as described above, we focus only on the submodel for the minority group that received stereotype primes, which is pictured in Figure 6. In the figure, paths with dashed lines and parameters in bold/italics signify group-specific parameters. These include the factor mean η, numerical factor loading λ_{num}, numerical intercept μ_{num}, and numerical unique variance ψ_{num}. Other parameters (not displayed) were restricted to be equal across all four groups, with the exception of the factor variance. The factor variance was restricted to be equal within both minority groups, and separately within both majority groups.

To carry out the tests, we first fit the four-group model to the data, calculating casewise derivatives and the observed information matrix. Second, to assess measurement invariance within the "minority, stereotype prime" group only, the scores of the n_{MSP} individuals from that group are ordered and aggregated along GPA (i.e., in this application the variable V). The cumulative score process (13) with respect to GPA then allows us to obtain various test statistics and p-values from this process. Test statistics include the double-max statistic from (15), the Cramér-von Mises statistic from (16), and the max LM statistic from (17). We focus on the double-max statistic due to its ease of interpretation and intuitive visual display.

As mentioned in the theory section, the tests give us the flexibility to study hypotheses of partial change: we have the ability to test various subsets of parameters. For the WDH data, we first test

$$H_0: \ (\lambda_{i,\text{num}} \ \eta_i \ \mu_{i,\text{num}} \ \psi_{i,\text{num}}) = (\lambda_{0,\text{num}} \ \eta_0 \ \mu_{0,\text{num}} \ \psi_{0,\text{num}}), \ i = 1, \ldots, n_{\text{MSP}}, \tag{25}$$

where $\lambda_{i,\text{num}}$ is the factor loading on the numerical scale, η_i is student i's factor mean, $\mu_{i,\text{num}}$ is student i's intercept on the numerical scale, and $\psi_{i,\text{num}}$ is student i's unique variance on the numerical scale. Thus, (25) states that these four group-specific model parameters are invariant with respect to GPA.

Results

The double-max test for the hypothesis (25) is shown in Figure 7, displaying the cumulative score processes for each of the four group-specific parameters across GPA along with horizontal lines representing the critical value for $\alpha = 0.05$. This shows that the "minority, stereotype prime" group lacks invariance with respect to GPA, because one of the processes crosses its boundaries. Furthermore, as this process pertains to the variance ψ_{num} of the numerical scale (bottom panel) it can be concluded that the invariance violation is associated with this parameter. Both the factor loading for the numerical scale λ_{num} and the factor mean η also exhibit increased fluctuation, although they are just non-significant at level $\alpha = 0.05$. Only the process associated with the intercept μ_{num} shows moderate (and hence clearly non-significant) fluctuation across GPA.

We can also use the test results to locate the threshold ν, which in this context can be used to define GPA-based subgroups whose measurement parameters differ. As described previously, information about the location of ν can be obtained by examining the peaks in Figure 7. The main peak occurs around a GPA of 6.3 and a second one near 5.9 (6 corresponds to the American 'C' range). These peaks roughly divide individuals into those receiving D's and F's, those receiving solid C's, and those receiving high C's, B's, and A's.

In addition to the double-max test, the Cramér-von Mises or max LM statistics could also be employed. They also lead to clearly significant violations of measurement invariance ($p < 0.005$ for both tests) and the corresponding visualizations bring out similar peaks as the double-max test in Figure 7 (and hence are omitted here).

Summary

Application of the proposed measurement invariance tests to the Wicherts et al. (2005) data allowed us to study the extent to which model parameters are invariant with respect to GPA in a straightforward manner. The tests focused on a set of group-specific parameters within a four-group confirmatory factor model, and they could be carried out using the results from a single estimated model. The latter fact is important, as other approaches to the problem described here may require multiple models to be estimated (e.g., for LR tests) or adversely impact model fit and degrees of freedom (e.g., if GPA is inserted directly into the model). Further, through examination of both test statistics and cumulative score processes, the tests were interpretable from both a theoretical and applied standpoint.

We included a factor mean in the application in order to demonstrate the tests' generality. To deal with measurement invariance specifically, however, we may elect to focus on only the measurement parameters within the model (omitting the factor mean). That is, it would not have been notable or surprising if we had observed that the factor mean lacked measurement invariance with respect to GPA: this result would have implied that individuals' latent intelligence fluctuates with GPA. We speculate that such a result would often be obtained when V is related to the latent variable of interest.

General Discussion

In this paper, we have applied a family of stochastic process-based statistical tests to the study of measurement invariance in psychometrics. The tests have reasonable power, can isolate subgroups of individuals violating measurement invariance based on a continuous auxiliary variable, and can isolate specific model parameters affected by the violation. In this section, we consider the tests' use in practice and their extension to more complex scenarios.

Use in Practice

The proposed tests give researchers a new set of tools for studying measurement invariance, allowing them the flexibility to: (1) simultaneously test all model parameters across all individuals, yielding results relevant to many types of measurement invariance (see, e.g., Meredith, 1993), (2) test a subset of model parameters, either across all individuals or within a multiple-group model, and (3) use the tests as a type of modification index following significant LR tests. Traditional steps to studying measurement invariance have involved sequential LR tests for various types of invariance using multiple-group models. As shown in the "Application" section, it can be beneficial to employ the traditional steps in tandem with the proposed tests. Furthermore, when groups are not defined in advance (e.g., continuous V), the traditional steps fail unless further assumptions about the nature of the groups are made.

It is worth mentioning that the proposed tests are not invariant with respect to the choice of identification constraints. They share this property with both LM and Wald tests (while LR tests are invariant to the choice of identification constraints). However, asymptotically the influence of the parametrization disappears and also in finite samples it typically does not change the results significantly. For example, in the stereotype threat application, we assessed the p-values of all three tests under the constraint that the abstract reasoning loading is fixed to 1 (see Figure 6) and compared them to the p-values under the constraint that the verbal loading is fixed to 1. All p-values remained clearly significant and the changes were all smaller than 0.00002.

In addition to providing general information about whether or not measurement invariance holds, the proposed tests allow researchers to interpret the nature of the invariance violation. This is made possible, e.g., through the tests' abilities to locate ν, the threshold dividing individuals into subgroups that violate measurement invariance (see Equation (10)). It is also possible to formally estimate one or more ν thresholds by adopting a (partially) segmented model (see e.g., Zeileis et al., 2010).

Comparison to Other Methods

The step functions from (10) were employed in the current paper to highlight connections with nested model comparison, but the proposed tests typically have non-trivial power for all (non-constant) deviation patterns $\theta(v_i)$ (considered in Equation (9)). Thus, the tests will also have power for linear deviations from parameter constancy, although other techniques may perform better in this particular case. Examples include the class of moderated factor models (Bauer & Hussong, 2009; Molenaar, Dolan, Wicherts, & van der Mass, 2010; Neale, Aggen, Maes, Kubarych, & Schmitt, 2006; Purcell, 2002), whereby continuous moderators are allowed to linearly impact model parameters (Purcell, 2002, also considers quadratic effects of the moderators). Moreover, if the change in parameters is continuous but not exactly linear – e.g., a sigmoidal shift from one set of parameters to another one – then the simple single shift model in (10) may be a useful first approximation and may have better power than linear techniques (depending on how clearly the two regimes are separated).

The proposed tests are also similar to factor mixture models (e.g., Dolan & van der Maas, 1998; Lubke & Muthén, 2005) in that they can handle unknown subgroups violating measurement invariance. Covariates can also be used to predict the unknown subgroups under both approaches. For example, for the data used in our simulations, one could employ a factor mixture model with two latent classes where the class probabilities depend on student's age through a logit link. If the assumptions of such a mixture hold, i.e., that each observation is a weighted combination from a (known) number of classes, the fitted model has more power uncovering this structure and makes it easy to interpret. However, the proposed tests are easier to use with fewer assumptions about the type of deviations (as long as they occur along the selected variable V).

More generally, all the approaches described above can be distinguished from the proposed tests in that they require estimation of a new model of increased complexity (due to the moderator/covariate). The proposed tests, on the other hand, are of the "posthoc" variety, relying only on results calculated during the original model estimation. While no method clearly dominates across all situations, use of the proposed tests can at least reduce some of the technical issues associated with estimating and interpreting models of greater complexity. This alone is not a good reason to use the tests, but it is a consideration that is often meaningful in practice.

Categorical Auxiliary Variable

One issue that was largely unaddressed in this paper involved the use of categorical V to study measurement invariance. In this case, groups are already specified in advance, and so traditional methods for fixed subgroups (i.e., LR, Wald, and LM tests) may suffice. However, we can also obtain an LM-type statistic from the framework developed here. Assume the observations are divided into C categories $I_1, I_2, \ldots, I_C$. Then, the increment of the cumulative score process $\Delta_{I_c}\boldsymbol{B}(\hat{\boldsymbol{\theta}})$ within each category is just the sum of the corresponding scores. In somewhat sloppy notation:

$$\Delta_{I_c}\boldsymbol{B}(\hat{\boldsymbol{\theta}}) \;=\; \hat{\boldsymbol{I}}^{-1/2} n^{-1/2} \sum_{i \in I_c} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; x_i). \qquad (26)$$

This results in a $C \times k$ matrix, with one entry for each category-by-model parameter combination. We can test a specific model parameter for invariance by focusing on the associated column of the $C \times k$ matrix and employing a weighted squared sum of the entries in the column to obtain a χ^2-distributed statistic with $(C - 1)$ degrees of freedom (Hjort & Koning, 2002). Alternatively, to simultaneously test multiple parameters, we can sum the χ^2 statistics and degrees of freedom for the individual parameters. In addition to categorical V, this framework may be useful for ordinal V (or continuous V with many ties). In this situation, one could also adapt the statistics (15), (16), and (17) computed from the same cumulative score process as usual. The only required modification is that the statistics should not depend on the process's values "within" a category (or group of ties). This is easily achieved by taking the maximum (or sum) not over all observations $i = 1, \ldots, n$, but over only those i at the "end" of a category.

The number of potential thresholds may be very low in these situations, which impacts the extent to which asymptotic results hold for the main test statistics described in this paper.

Extensions

The family of tests described in this paper can be extended in various ways. First, it is possible to construct an algorithm that recursively defines groups of individuals violating measurement invariance with respect to multiple auxiliary variables. Such an algorithm is related to classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984; Merkle & Shaffer, 2011; Strobl, Malley, & Tutz, 2009), with related algorithms being developed for general parametric models (Zeileis, Hothorn, & Hornik, 2008) and Rasch models in particular (Strobl, Kopf, & Zeileis, 2010).

Relatedly, Sánchez (2009) describes a general method for partitioning/segmenting structural equation models within a partial least squares framework. This method involves direct maximization of the likelihood ratio (i.e., fitting the model for various subgroups defined by V and choosing the subgroups with the largest likelihood ratio). Thus, unlike the tests described in this paper, this approach does not provide a formal significance test with a controlled level of type I errors.

The proposed tests also readily extend to other psychometric models. For example, the tests can be used to generally study the stability of structural equation model parameters across observations. This includes the study of measurement invariance in second-order growth models and related models for longitudinal data (e.g., Ferrer, Balluerka, & Widaman, 2008; McArdle, 2009). The issue of measurement invariance is important in these models in order to establish that a “true” change has occurred in the individuals to which the model has been fitted. The tests described here can be used to help establish measurement invariance with respect to both continuous and categorical auxiliary variables, using only output from the estimated model of interest.

Finally, the tests can be used for the study of differential item functioning (DIF) in item response models (e.g., Strobl et al., 2010, who focused on Rasch models). Traditional DIF methods are similar to those for factor analysis in that subgroups must be specified in advance. The tests proposed here can detect subgroups automatically, however, offering a unified way of studying both measurement invariance in factor analysis and differential item functioning in item response. While factor-analytic measurement invariance methods and DIF methods have developed largely independently of one another, the methods can certainly be treated from a unified perspective (e.g., McDonald, 1999; Millsap, 2011; Stark, Chernyshenko, & Drasgow, 2006). The tests proposed here were designed with this perspective in mind.

Summary

We outline a family of stochastic process-based parameter instability tests from theoretical statistics and apply them to the issue of measurement invariance in psychometrics. The paper includes theoretical development, an applied example, and study of the tests’ performance under controlled conditions. The tests are found to have good properties via simulation, making them useful for many psychometric applications. More generally, the tests help solve standing problems in measurement invariance research and provide many avenues for future research. This can happen both through extensions of the tests within a factor-analytic context and through application of the tests to new models.

Computational Details

All results were obtained using the R system for statistical computing (R Development Core Team, 2012), version 3.0.3, employing the add-on packages lavaan 0.5-15 (Rosseel, 2012) and OpenMx 1.3.2-2301 (Boker et al., 2011) for fitting of the factor analysis models and strucchange 1.5-0 (Zeileis, Leisch, Hornik, & Kleiber, 2002; Zeileis, 2006) 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/` while OpenMx is available under the Apache License 2.0 from `http://OpenMx.psyc.virginia.edu/`. R code for replication of our results is available at `http://semtools.R-Forge.R-project.org/`.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, *61*, 821–856.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3–29.

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.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306–317. doi: 10.1007/S11336-010-9200-6

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley & Sons.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), S176–S181.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society B*, *37*, 149–163.

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

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

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, *4*, 22–36.

Hansen, B. E. (1992). Testing for parameter instability in linear models. *Journal of Policy Modeling*, *14*, 517–533.

Hansen, B. E. (1997). Approximate asymptotic *p* values for structural-change tests. *Journal of Business & Economic Statistics*, *15*, 60–67.

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

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.

Merkle, E. C., & Shaffer, V. A. (2011). Binary recursive partitioning methods with application to psychology. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 161–181.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153–171). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Neale, M. C., Aggen, S. H., Maes, H. H., Kubarych, T. S., & Schmitt, J. E. (2006). Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors*, *31*, 1010–1034.

Nyblom, J. (1989). Testing for the constancy of parameters over time. *Journal of the American Statistical Association*, *84*, 223–230.

Ploberger, W., & Krämer, W. (1992). The CUSUM test with OLS residuals. *Econometrica*, *60*(2), 271–285.

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/`

Sánchez, G. (2009). *PATHMOX approach: Segmentation trees in partial least squares path modeling* (Unpublished doctoral dissertation). Universitat Politécnica de Catalunya.

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

Shorack, G. R., & Wellner, J. A. (1986). *Empirical processes with applications to statistics.* New York: John Wiley & Sons.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306.

Strobl, C., Kopf, J., & Zeileis, A. (2010, December). *A new method for detecting differential item functioning in the Rasch model* (Technical Report No. 92). Department of Statistics, Ludwig-Maximilians-Universität München. URL `http://epub.ub.uni-muenchen.de/11915/`.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323–348.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–716.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples.* Mahwah, NJ: Lawrence Erlbaum Associates.

Zeileis, A. (2005). A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, *24*(4), 445–466.

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

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

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*, 492–514.

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, *7*(2), 1–38. URL `http://www.jstatsoft.org/v07/i02/`.

Zeileis, A., Shah, A., & Patnaik, I. (2010). Testing, monitoring, and dating structural changes in exchange rate regimes. *Computational Statistics & Data Analysis*, *54*, 1696–1706.

Author Note

This work was supported by National Science Foundation grant SES-1061334. The authors thank Jelte Wicherts, who generously shared data for the stereotype threat application, Yves Rosseel, who provided feedback and code for performing the tests with the lavaan package, Kris Preacher, who provided helpful comments on the manuscript, and the participants of the Psychoco 2012 workshop on psychometric computing for helpful discussion. Correspondence to Edgar C. Merkle, Department of Psychological Sciences, University of Missouri, Columbia, MO 65211. Email: merklee@missouri.edu.

Table 1

Simulated power for three test statistics across four sample sizes n, nine magnitudes of measurement invariance violations, and two subsets of tested parameters k^*. See Figure 5 for a visualization (using all 17 violation magnitudes).

			Violation Magnitude (SE)								
n	k^*	Statistic	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
50	3	*DM*	2.1	2.9	5.8	10.0	15.5	21.9	24.6	21.7	16.0
		CvM	3.8	5.4	11.4	21.5	33.4	47.1	49.3	43.1	30.3
		$\max LM$	6.1	6.6	10.3	16.7	25.4	36.1	37.6	32.4	23.7
	19	*DM*	1.2	1.4	2.0	3.1	4.4	6.1	6.7	6.5	6.8
		CvM	3.4	3.7	5.5	8.1	10.9	15.6	18.6	18.9	17.9
		$\max LM$	7.7	8.1	8.7	10.9	13.0	16.0	16.7	17.1	15.9
100	3	*DM*	2.8	4.0	7.8	15.2	26.2	39.7	55.3	66.1	70.1
		CvM	4.5	6.7	12.1	26.1	44.1	63.0	79.1	87.2	90.1
		$\max LM$	5.2	6.8	10.8	19.8	34.0	51.8	69.9	80.9	85.7
	19	*DM*	2.4	2.4	3.4	5.6	8.8	14.6	22.1	29.9	34.3
		CvM	4.6	4.7	6.3	9.8	16.1	24.1	35.1	46.8	53.3
		$\max LM$	6.9	7.5	8.1	10.5	14.9	20.0	28.1	37.8	43.1
200	3	*DM*	3.5	4.4	8.5	17.5	30.7	50.3	68.4	82.3	91.9
		CvM	4.8	6.6	12.9	26.6	46.1	69.1	86.2	94.7	98.4
		$\max LM$	5.3	6.7	10.8	21.4	36.6	59.5	78.8	91.5	97.4
	19	*DM*	3.6	3.7	4.4	7.3	11.3	21.2	35.8	51.5	67.2
		CvM	4.4	4.8	7.3	11.5	18.7	30.1	46.6	62.6	76.4
		$\max LM$	6.3	6.4	7.4	10.6	15.5	23.9	37.9	53.4	69.2
500	3	*DM*	4.1	5.3	10.8	19.5	34.2	52.2	71.8	86.4	95.6
		CvM	5.2	6.2	13.6	27.3	47.7	68.7	86.3	95.6	99.0
		$\max LM$	5.7	6.1	10.9	21.0	39.2	60.5	81.1	92.9	98.5
	19	*DM*	4.2	4.0	5.7	8.8	14.5	25.7	42.8	62.1	78.8
		CvM	4.5	5.0	7.4	12.5	20.7	33.9	50.2	68.7	82.9
		$\max LM$	5.2	6.5	7.6	9.8	16.1	26.3	42.3	60.6	78.2

Abbreviations: CvM = Cramér-von Mises test; $\max LM$ = Maximum Lagrange multiplier test; DM = Double-max test.

Figure Captions

Figure 1. Path diagram representing the base factor analysis model used for the artificial example and simulations. To induce measurement invariance violations, a seventh observed variable (student age) determines the values of the verbal factor loadings (λ_{11}, λ_{21}, λ_{31}).

Figure 2. A comparison of the $\max LM$ (solid line) and $\max LR$ (dashed line) test statistics for (8) (i.e., $k^* = 19$). The gray horizontal line corresponds to the critical value at $\alpha = 0.05$ while the dotted vertical line highlights the threshold at which both test statistics assume their maximum.

Figure 3. Cumulative score processes for each verbal factor loading. The solid gray horizontal lines correspond to the critical value of the double-max test at $\alpha = 0.05$.

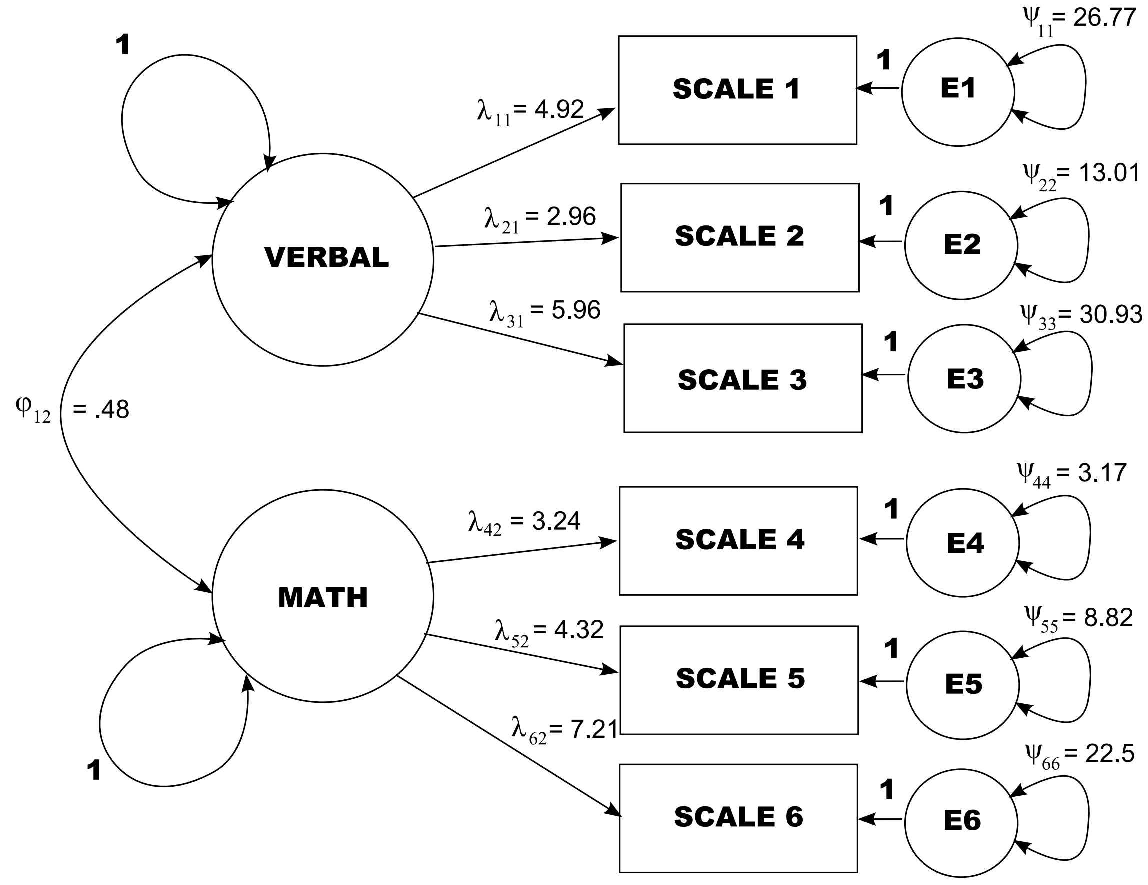
Figure 4. Three test statistics of (24) (with $k^* = 3$) and (8) (with $k^* = 19$), based on the example involving measurement invariance with respect to student age. Solid gray horizontal lines represent critical values at $\alpha = 0.05$, and the dotted, horizontal lines (second row) represent values of the Cramér-von Mises test statistic.

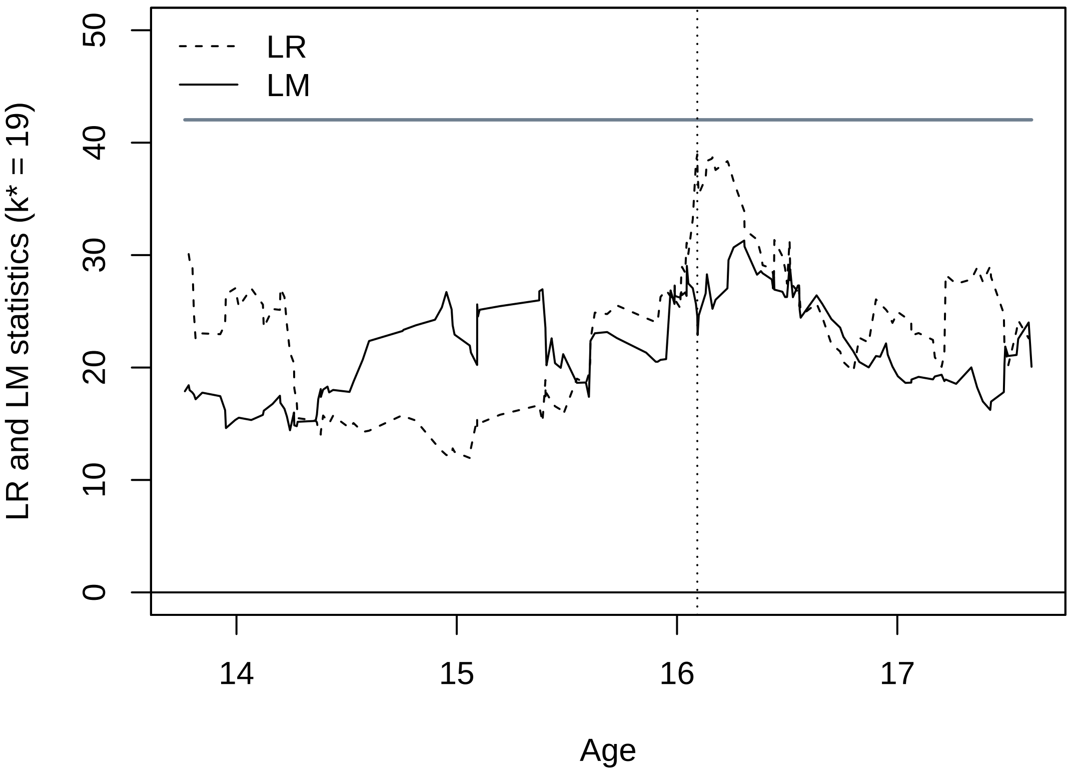
Figure 5. Simulated power curves for the double-max test (solid), Cramér-von Mises test (dashed), and $\max LM$ test (dotted) across four sample sizes n, two subsets of tested parameters k^*, and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$). See Table 1 for the underlying numeric values (using a subset of nine violation magnitudes).

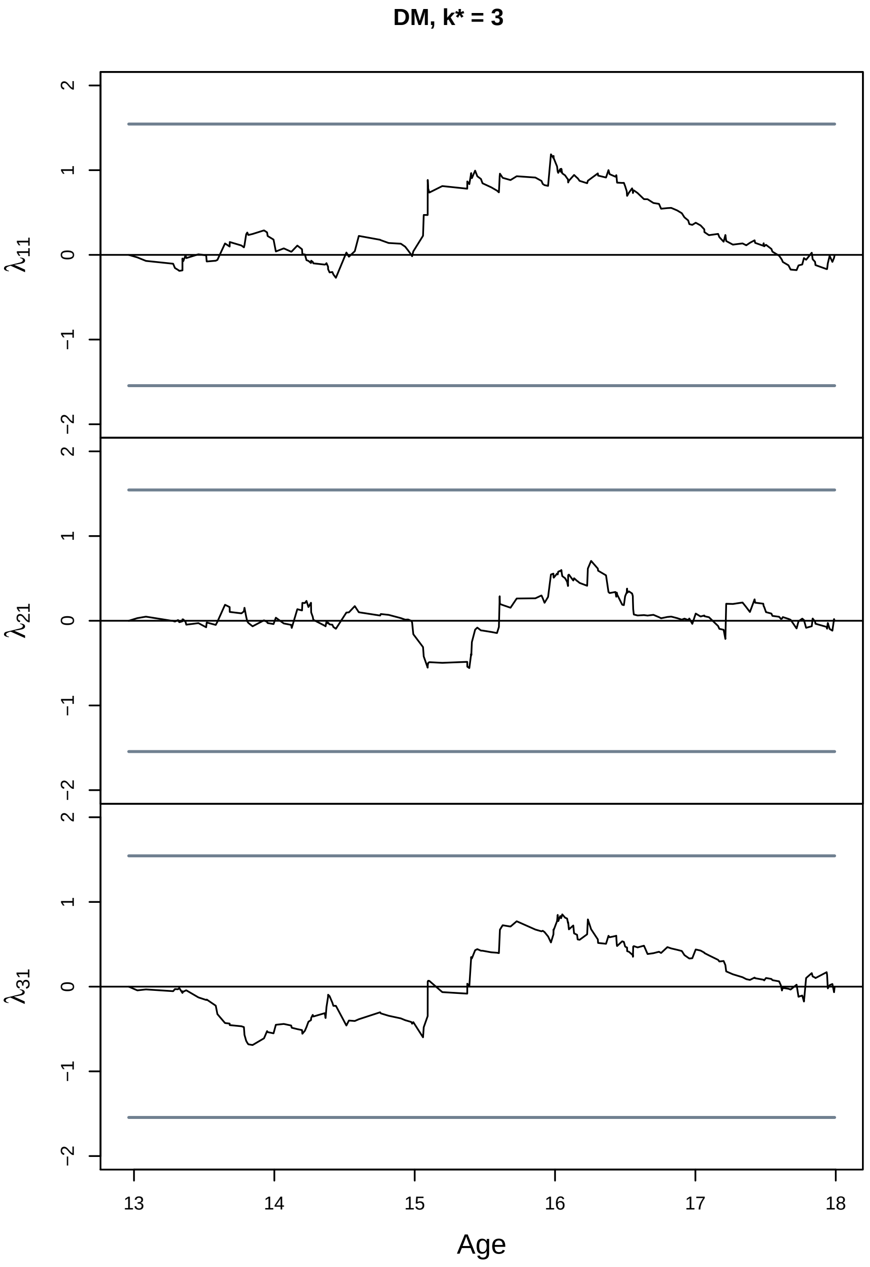
Figure 6. Path diagram representing the submodel for the “minority, stereotype prime” group in model 5b of WDH. Dashed paths and parameters in bold/italics are group specific. Other parameters (not displayed) were restricted to be equal across all four

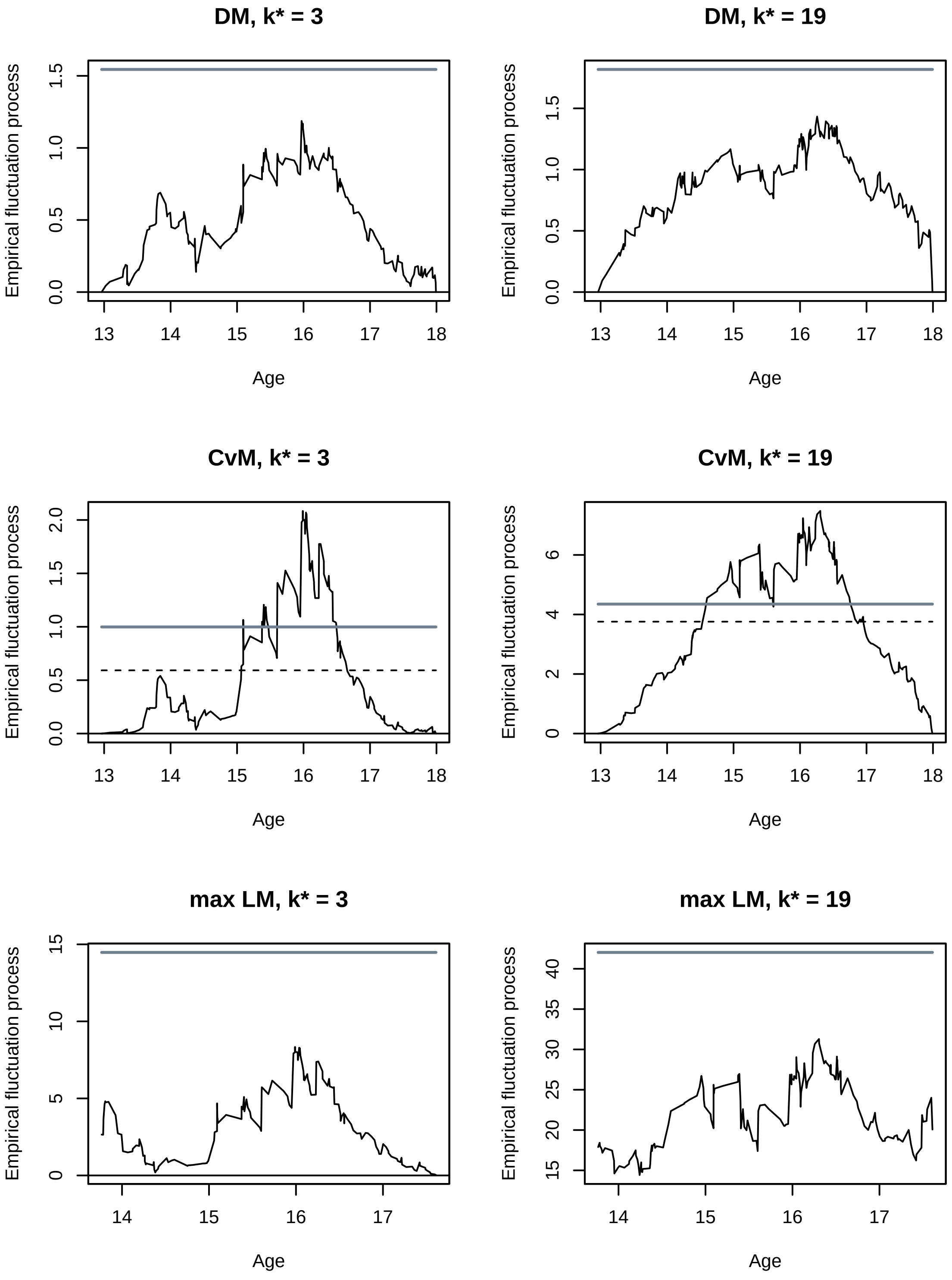
groups, with the factor variance only being restricted across the two minority groups.

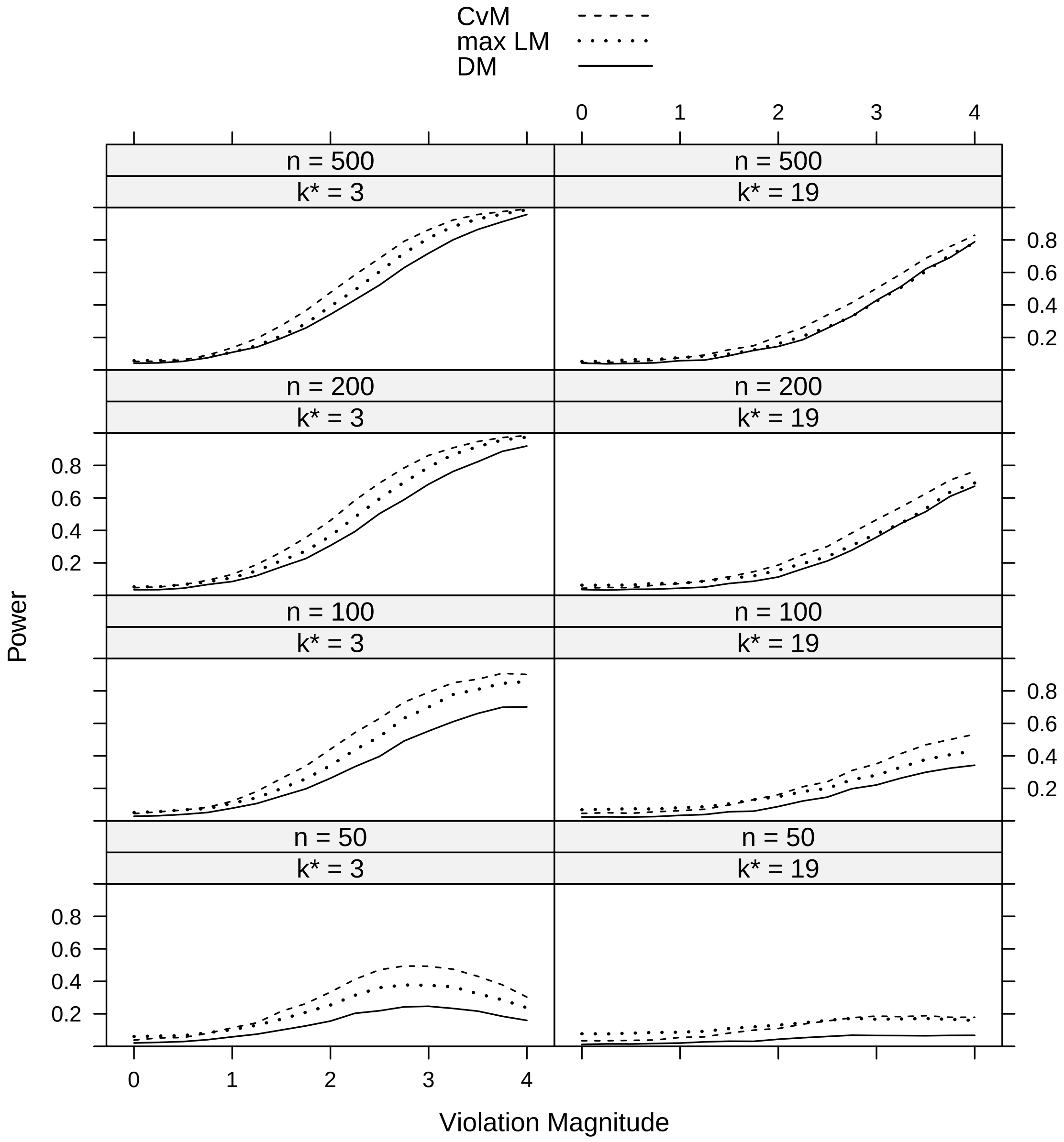
Figure 7. Cumulative score processes for the four group-specific parameters across GPA in the “minority, stereotype prime” group using data from Study 1 of WDH. The solid gray horizontal lines correspond to the critical value of the double-max test at $\alpha = 0.05$.

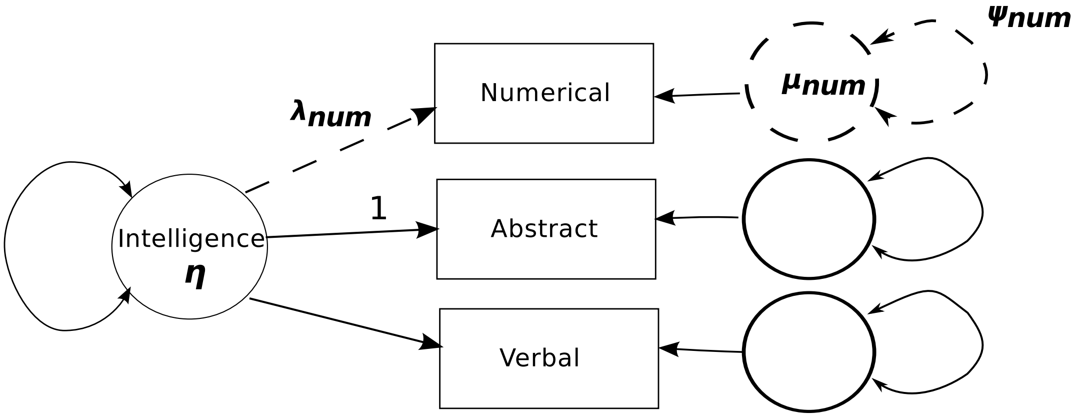

1
VERBAL
φ_{12} = .48
MATH
1
λ_{11}= 4.92
λ_{21} = 2.96
λ_{31} = 5.96
λ_{42} = 3.24
λ_{52} = 4.32
λ_{62}= 7.21
SCALE 1
SCALE 2
SCALE 3
SCALE 4
SCALE 5
SCALE 6
1
1
1
1
1
1
E1
E2
E3
E4
E5
E6
ψ_{11}= 26.77
ψ_{22}= 13.01
ψ_{33}= 30.93
ψ_{44} = 3.17
ψ_{55} = 8.82
ψ_{66} = 22.5


LR
LM
LR and LM statistics (k* = 19)
0
10
20
30
40
50
14
15
16
17
Age

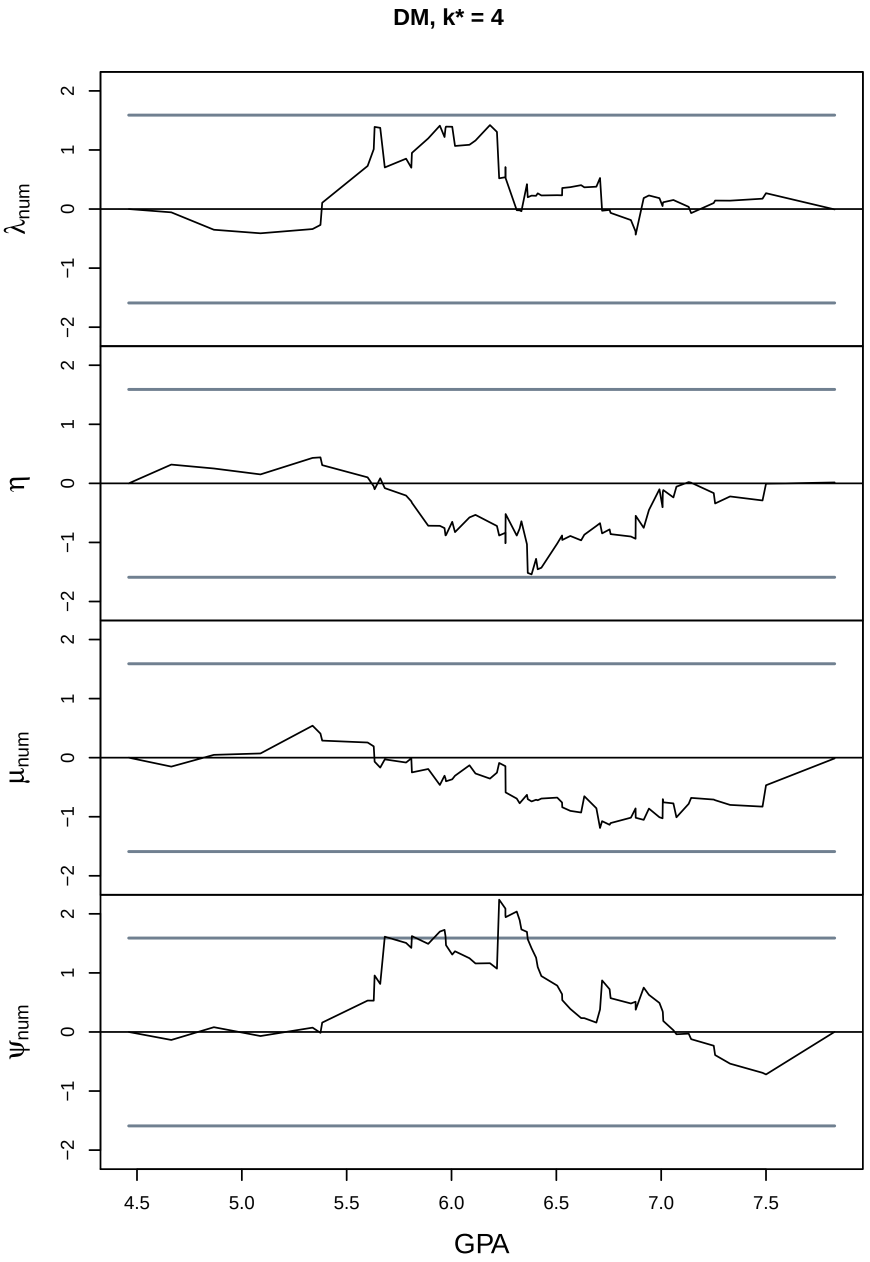

DM, k* = 3
λ_{11}
λ_{21}
λ_{31}
2
1
0
−1
−2
13
14
15
16
17
18
Age


DM, k* = 3
DM, k* = 19
CvM, k* = 3
CvM, k* = 19
max LM, k* = 3
max LM, k* = 19
Empirical fluctuation process
Age
0.0
0.5
1.0
1.5
2.0
0
2
4
6
5
10
15
20
25
30
35
40
13
14
15
16
17
18


CvM
max LM
DM
n = 500
k* = 3
n = 500
k* = 19
n = 200
k* = 3
n = 200
k* = 19
n = 100
k* = 3
n = 100
k* = 19
n = 50
k* = 3
n = 50
k* = 19
Power
0.2
0.4
0.6
0.8
0
1
2
3
4
Violation Magnitude



ψnum
μnum
Numerical
λnum
1
Abstract
Intelligence
η
Verbal


DM, k* = 4
λnum
η
μnum
ψnum
2
1
0
−1
−2
4.5
5.0
5.5
6.0
6.5
7.0
7.5
GPA