

Generalized Measurement Invariance Tests for Factor Analysis

Ed Merkle¹ Achim Zeileis²

¹University of Missouri

²Universität Innsbruck

Supported by grant SES-1061334 from the U.S. National
Science Foundation

1 Background

2 Proposed Tests

3 Illustration

4 Conclusions

Measurement Invariance

- Measurement invariance: Sets of tests/items consistently assigning scores across diverse groups of individuals.
- Notable violations of measurement invariance:
 - SAT for different ethnic groups (Atkinson, 2001)
 - Intelligence tests & the Flynn effect (Wicherts et al., 2004)

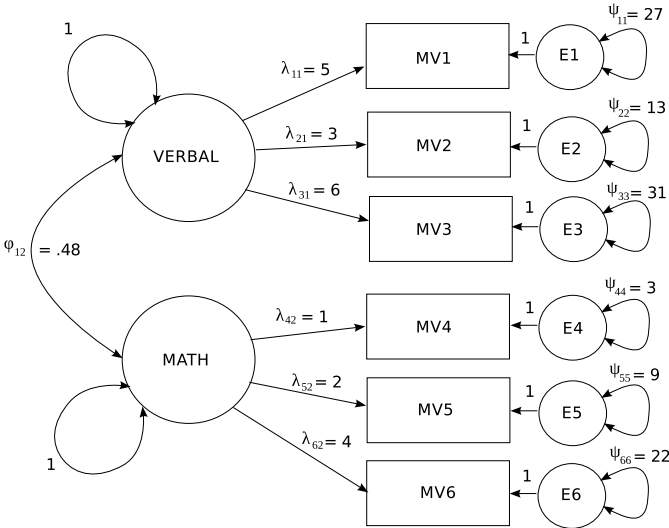
Example (Age ≤ 16)

Background

Proposed
Tests

Illustration

Conclusions



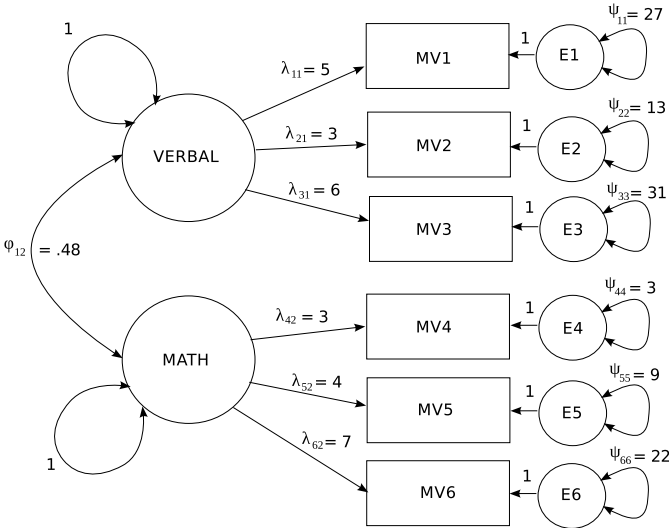
Example (Age > 16)

Background

Proposed
Tests

Illustration

Conclusions



Hypotheses

- Hypothesis of “full” measurement invariance:

$$H_0 : \boldsymbol{\theta}_i = \boldsymbol{\theta}_0, i = 1, \dots, n$$

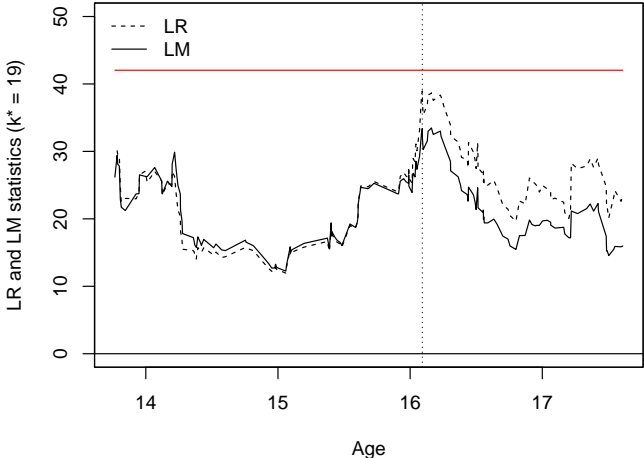
$$H_1 : \text{Not all the } \boldsymbol{\theta}_i = \boldsymbol{\theta}_0$$

where $\boldsymbol{\theta}_i = (\lambda_{i,1,1}, \dots, \psi_{i,1,1}, \dots, \varphi_{i,1,2})^\top$ is the full p -dimensional parameter vector for individual i .

Hypotheses

- H_0 from the previous slide is difficult to fully assess due to all the ways by which individuals may differ.
- We typically place people into groups based on a meaningful auxiliary variable, then study measurement invariance across those groups (via Likelihood Ratio tests, Lagrange multiplier tests, Wald tests).
- If we did not know the groups in advance, we could conduct a LR or LM test for each possible grouping, then take the maximum. Requires different critical values! (Can be obtained from proposed tests.)

Lack of Grouping



Proposed Tests

- In contrast to existing tests of measurement invariance, the proposed tests offer the abilities to:
 - Test for measurement invariance when groups are ill-defined (e.g., when the grouping variable is continuous).
 - Test for measurement invariance in any subset of model parameters.
 - Interpret the nature of measurement invariance violations.

Proposed Tests

- The proposed family of tests rely on first derivatives of the model's log-likelihood function.
- We consider individual terms (*scores*) of the gradient. These scores tell us how well a particular parameter describes a particular individual.

$$\sum_{i=1}^n s(\hat{\theta}; \mathbf{x}_i) = \mathbf{0}, \text{ where}$$

$$s(\hat{\theta}; \mathbf{x}_i) = \frac{\partial}{\partial \theta} \log L(\mathbf{x}_i, \theta) \Big|_{\theta=\hat{\theta}}$$

Proposed Tests

- Under measurement invariance, parameter estimates should roughly describe everyone equally well. So people's scores should fluctuate around zero.
- If measurement invariance is violated, the scores should stray from zero.

Aggregating Scores

- We need a way to aggregate scores across people so that we can draw some general conclusions.
 - Order individuals by an auxiliary variable.
 - Define $t \in (1/n, n)$. The *empirical cumulative score process* is defined by:

$$\mathbf{B}(\hat{\theta}; t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} s(\hat{\theta}; \mathbf{x}_i).$$

where $\lfloor nt \rfloor$ is the integer part of nt .

- Under the hypothesis of measurement invariance, a functional central limit theorem holds:

$$\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1/2} \mathbf{B}(\hat{\boldsymbol{\theta}}; \cdot) \xrightarrow{d} \mathbf{B}^0(\cdot),$$

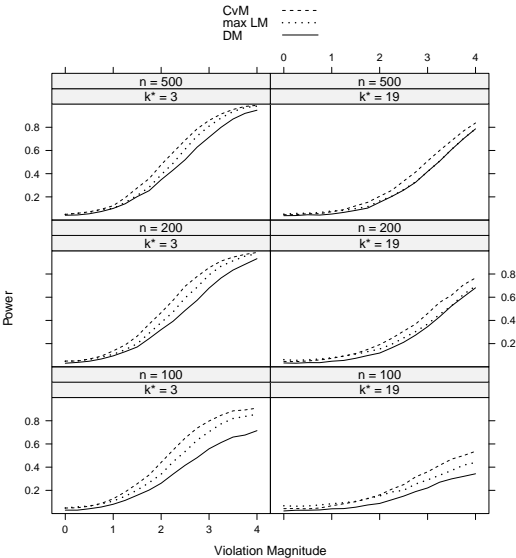
where $\mathbf{I}(\hat{\boldsymbol{\theta}})$ is the observed information matrix and $\mathbf{B}^0(\cdot)$ is a p -dimensional Brownian bridge.

- Testing procedure: Compute an aggregated statistic of the empirical score process and compare with corresponding quantile of aggregated Brownian motion.
- Test statistics: Special cases include double maximum (DM), Cramér-von Mises (CvM), maximum of LM statistics.

Simulation

- Simulation: What is the power of the proposed tests?
 - Two-factor model, with three indicators each.
 - Measurement invariance violation in three factor loading parameters, with magnitude from 0–4 standard errors.
 - Sample size in $\{100, 200, 500\}$.
 - Model parameters tested in $\{3, 19\}$.
 - Three test statistics.

Simulation



Example

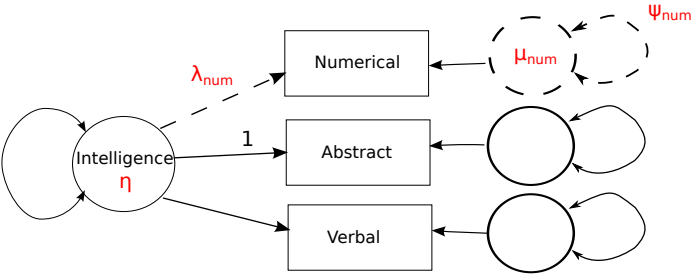
- Example: Studying stereotype threat via factor analysis (Wicherts et al., 2005)
 - Stereotype threat: Knowledge of stereotypes about one's social group might cause one to fulfill the stereotypes.
 - Wicherts et al. study: 295 students were administered three intelligence tests. Stereotypes were primed for half of the students.
 - Groups defined by: Ethnicity (majority/minority) and whether or not stereotypes were primed.

Model

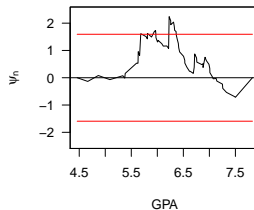
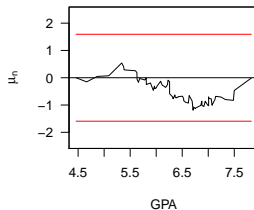
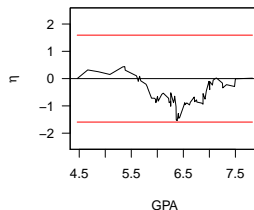
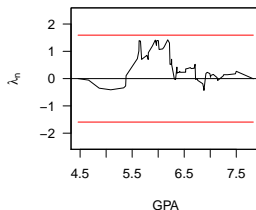
- To study the data, Wicherts et al. employed a series of four-group, one-factor models.
 - General finding: Minorities with stereotype primes have different measurement parameters than other groups.
 - Current example: Is measurement further impacted by academic performance (as measured by student GPA)?

Model

- We utilize a model employed by Wicherts et al., where four model parameters are specific to the “minority, stereotype prime” group.
 - Test for measurement invariance in these parameters wrt the student GPA variable (either all four together or individually).
 - Violations of measurement invariance imply that stereotype threat is more problematic for students of low or high GPA.

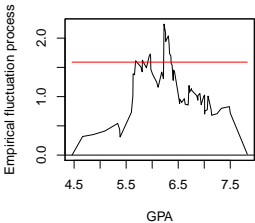


Results for Single Parameters

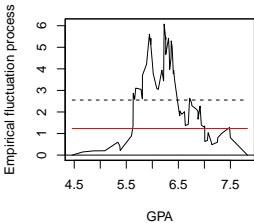


Aggregated Results

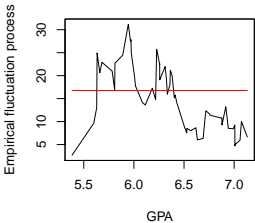
Aggregated Process, Double Max



Aggregated Process, CvM



Aggregated Process, max LM



Conclusions

- Measurement invariance tests utilizing stochastic processes have important advantages over existing tests:
 - Isolating specific parameters that violate measurement invariance, allowing the researcher to define specific types of measurement invariance “post hoc” instead of “a priori” .
 - Isolating groups of individuals whose parameter values differ.
 - Studying the impact of continuous variables on model estimates, without “ruining” the rest of the model.
- Power is reasonable, with specific tests being better in specific circumstances.

Software

- To carry out the tests, we utilize
 - `lavaan` for model estimation.
 - `estfun()` for score extraction, which is currently a combination of our own code and `lavaan` code.
 - `strucchange` for carrying out the proposed tests with the scores.
 - Required input: Fitted model, function for score extraction, and information matrix (optional).
 - `gefp()` constructs the process.
 - `sctest()` and `plot()` calculate and visualize test statistics.

Current Work

- Continued test implementation via `strucchange` and `lavaan` (and possibly `OpenMx`).
- Detailed examination of test properties.
- Extension to related psychometric issues.
- Working paper:
<http://econpapers.repec.org/RePEc:inn:wpaper:2011-09>

- Questions?