Generalized Measurement Invariance Tests for Factor Analysis

Ed Merkle[1] Achim Zeileis[2]

[1]University of Missouri

[2]Universität Innsbruck

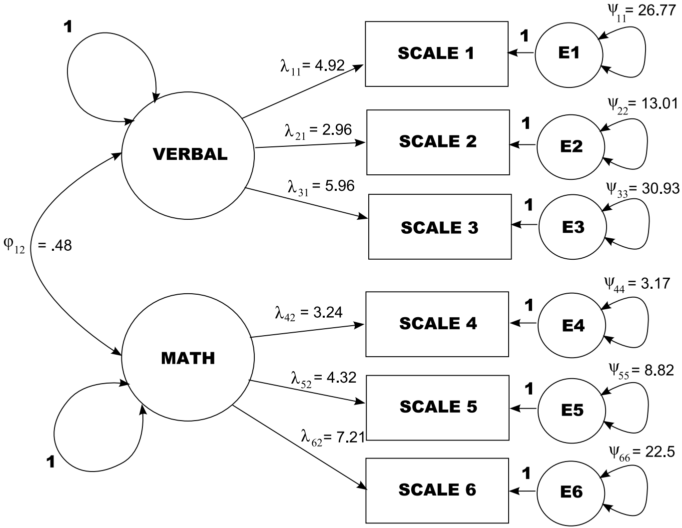
Supported by grant SES-1061334 from the U.S. National Science Foundation

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)

Studying Measurement Invariance



Studying Measurement Invariance

- Hypothesis of “full” measurement invariance:

$$H_0 : \boldsymbol{\theta}_i = \boldsymbol{\theta}_0, i = 1, \ldots, n$$
$$H_1 : \text{Not all the } \boldsymbol{\theta}_i = \boldsymbol{\theta}_0$$

Studying Measurement Invariance

- 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).

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.

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

$$\psi(\mathbf{x}_i, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathrm{L}(\mathbf{x}_i, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

Proposed Tests

- We can also consider individual terms (*scores*) of the gradient. These scores tell us how well a particular parameter describes a particular individual. If your score is 0 for some parameter, that parameter describes you well. If your score is far from 0, that parameter describes you poorly.

$$\psi(\mathbf{x}_i, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathrm{L}(\mathbf{x}_i, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

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:

$$\mathbf{W}_n(t, \boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \psi(\mathbf{x}_i, \boldsymbol{\theta}).$$

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

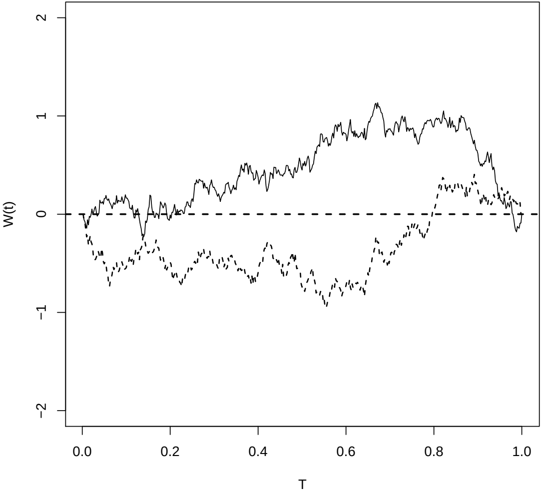
Theorem

- Theorem (Zeileis & Hornik, 2007): Under the hypothesis of measurement invariance, the following functional central limit theorem holds:

$$\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1/2}\mathbf{W}_n(\cdot, \widehat{\boldsymbol{\theta}}) \xrightarrow{d} \mathbf{W}^0(\cdot),$$

where $\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1/2}$ is the observed information matrix and $\mathbf{W}^0(\cdot)$ is a p-dimensional Brownian bridge.

Brownian Bridge



Software

- To carry out the tests, we use
  - `lavaan` for model estimation.
  - A combination of `lavaan` and our own code for calculating scores of fitted models.
  - `strucchange` for carrying out the proposed tests with the scores.

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: XX students administered three intelligence tests. Stereotypes were primed for one group of students.
  - Groups defined by: Ethnicity (majority/minority) and whether or not stereotypes were primed.

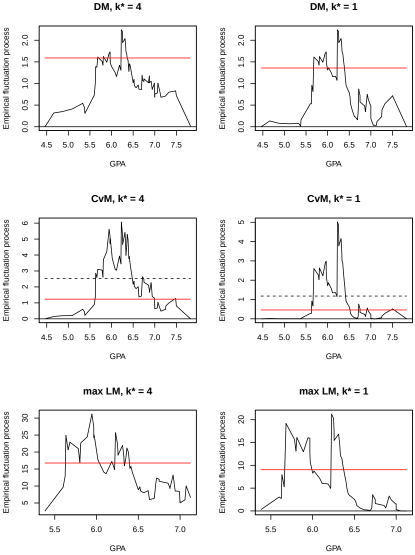
Model

- 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 only the factor mean).
  - Violations of measurement invariance imply that stereotype threat is more problematic for students of low (or high) GPA.

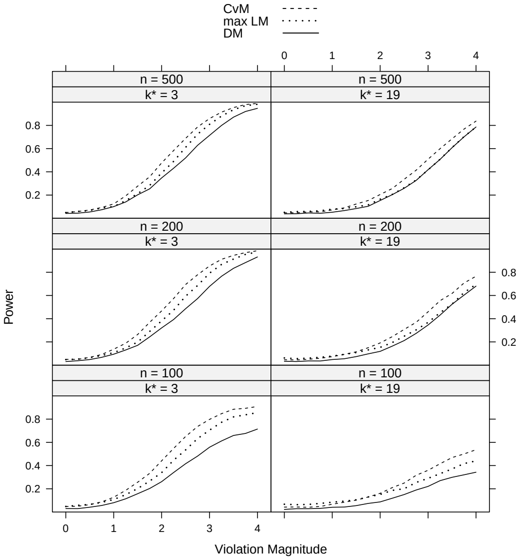
Results


DM, k* = 4
DM, k* = 1
CvM, k* = 4
CvM, k* = 1
max LM, k* = 4
max LM, k* = 1
Empirical fluctuation process
GPA

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


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
Power
Violation Magnitude
0
1
2
3
4
0.2
0.4
0.6
0.8

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.

Current Work

- Continued test implementation via `strucchange` and `lavaan` (and possibly `OpenMx`).
- Detailed examination of test properties via simulation.
- Extension to related psychometric issues (modification indices, mediation).

- Questions?