Testing for Measurement Invariance with respect to an Ordinal Variable

Ed Merkle[1] Jinyan Fan[2] Achim Zeileis[3]

[1]University of Missouri

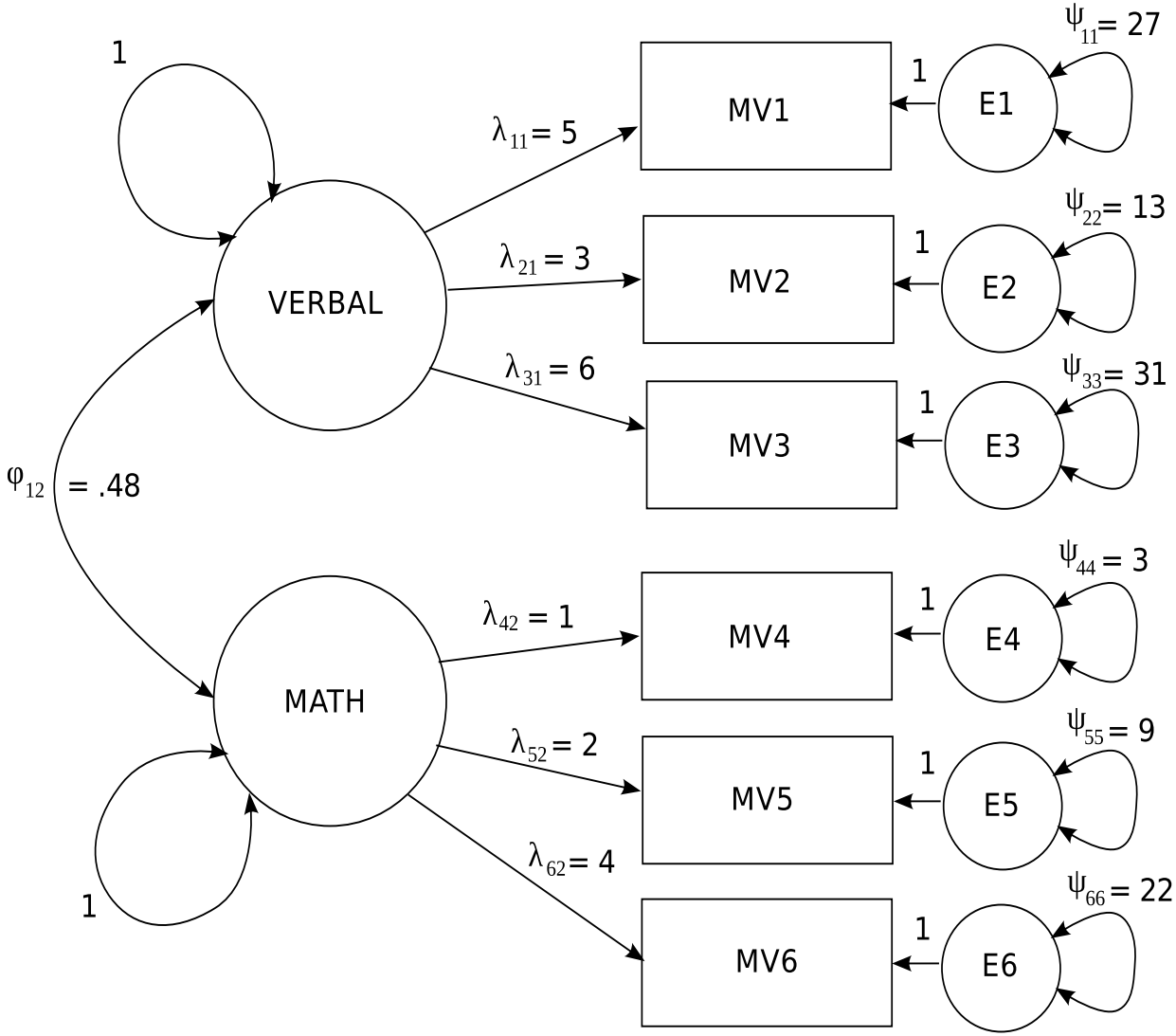
[2]Auburn University

[3]Universität Innsbruck

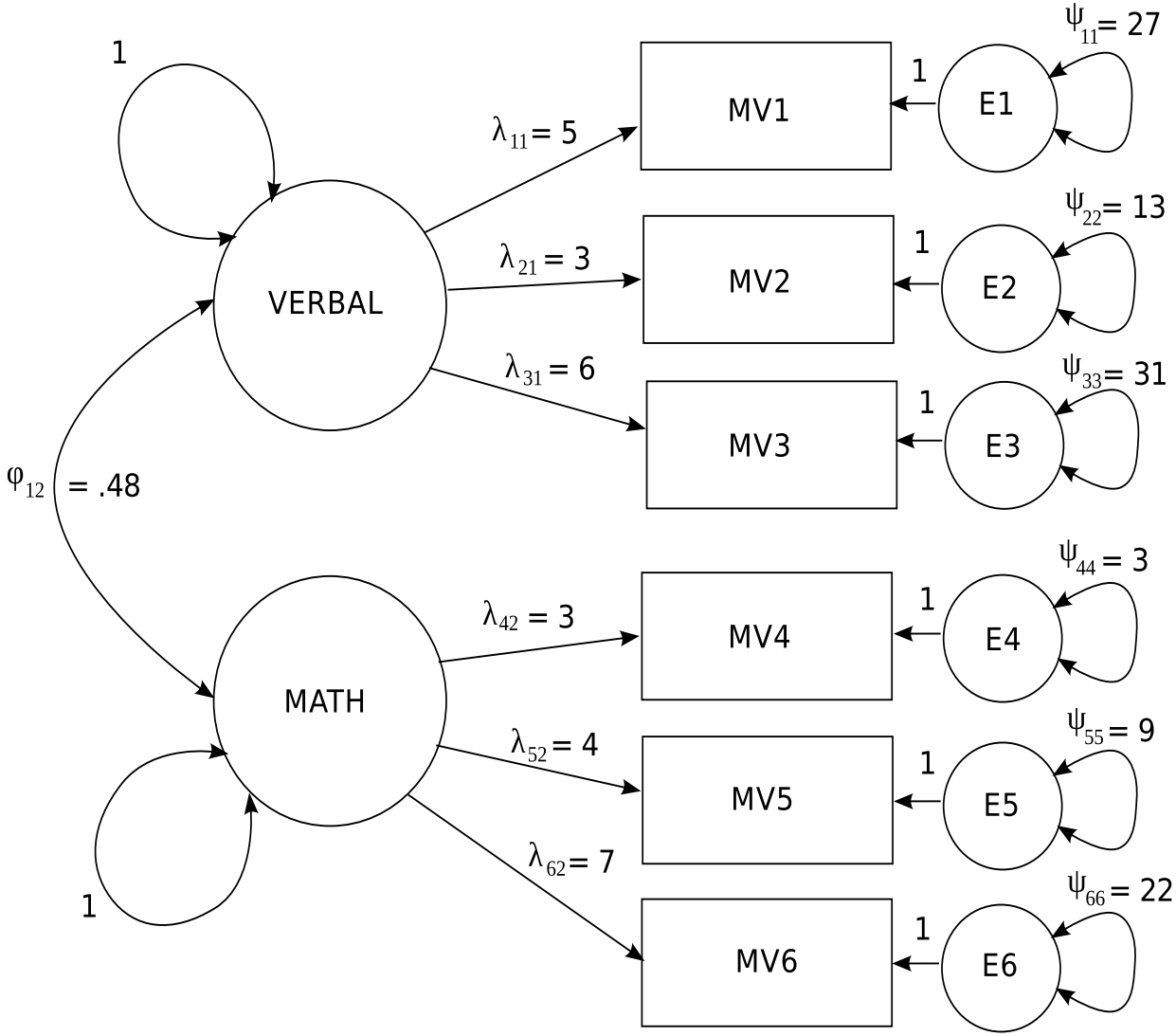
Measurement Invariance

- Measurement invariance: Sets of tests/items consistently assigning scores across diverse groups of individuals.
- Commonly studied via factor analysis (today's focus) and item response models.
- Focal problem: How do we test for measurement invariance with respect to an ordinal auxiliary variable, in a way that "respects" the fact that the variable is ordinal?

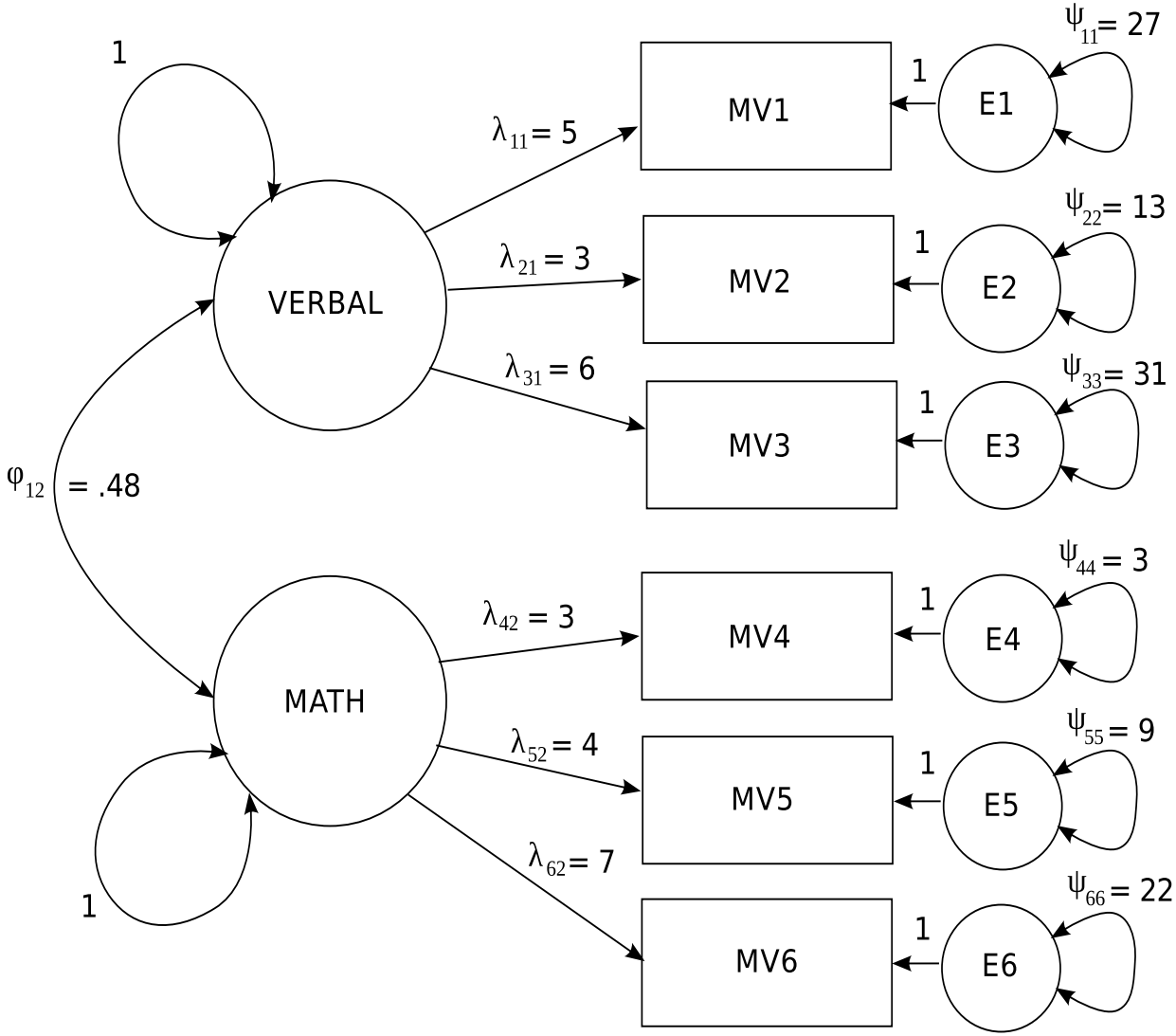
Example (13–14 years)



Example (15–16 years)



Example (17 years)



Outline

- The traditional way to test for measurement invariance here involves multiple-group models and likelihood ratio tests. However, these tests treat the ordinal variable as unordered, which can sometimes be suboptimal.

- In this talk, we:
  - Review last year's family of measurement invariance tests, which can be used when the auxiliary variable is continuous.
  - Present new test statistics to deal with situations when the auxiliary variable is ordinal.
  - Illustrate the statistics and compare them to traditional methods.

Hypotheses

- Hypothesis of "full" measurement invariance:

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

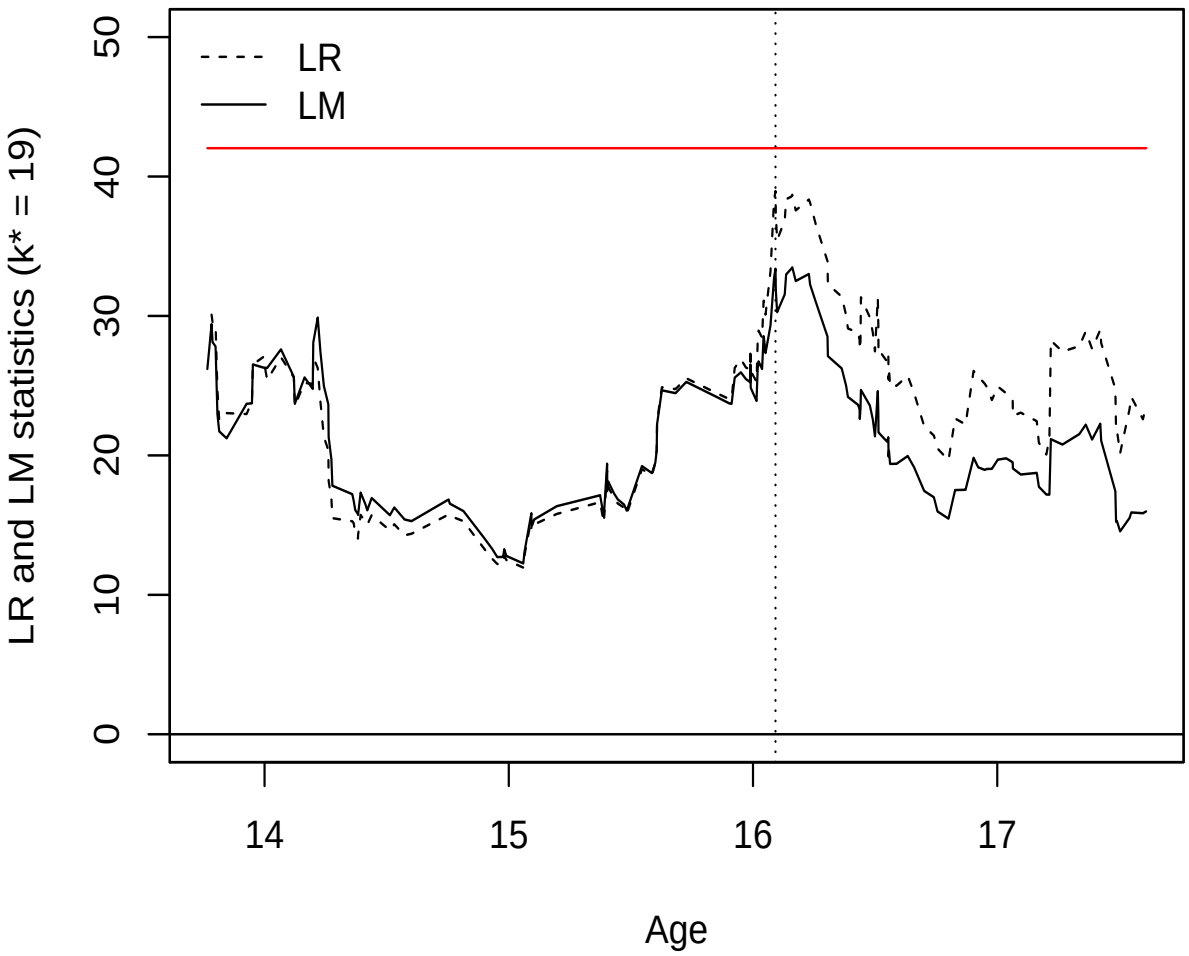
where $\boldsymbol{\theta}_i = (\lambda_{i,1,1}, \ldots, \psi_{i,1,1}, \ldots, \varphi_{i,1,2})^\top$ is the full p-dimensional parameter vector for individual i.

Hypotheses

- H_0 from the previous slide is difficult to fully assess due to all the ways by which individuals may differ.

- If our auxiliary variable of interest is continuous (i.e., we do not know the groups in advance), we could conduct a LR or LM test for each possible pair of groups, then take the maximum. Requires different critical values!

Lack of Grouping



Instability Tests

- The tests presented last year expand on this idea, removing the need to fit more than one model.

- Idea: Fit a "reduced" model to the data (assuming, e.g., that parameters are equal across groups). Examine how individuals' *scores* vary across the continuous auxiliary variable.

Scores

- What are scores?
  - The traditional way to obtain maximum likelihood estimates involves choosing estimates $\widehat{\boldsymbol{\theta}}$ so that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} \log \mathrm{L}(\mathbf{x}_i, \boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}$$

  - Individual terms of this summation are the scores

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

  - $s(\hat{\boldsymbol{\theta}}; \mathbf{x}_i)$ can be conceptualized as a residual for individual i: close to $\mathbf{0}$ is good, far from $\mathbf{0}$ is not so good.

Instability Tests

- Under measurement invariance, parameter estimates should roughly describe everyone equally well. So, people's scores should fluctuate around zero as we move up the auxiliary variable.

- If measurement invariance is violated, the scores should stray from zero around specific points of the auxiliary variable.

Aggregating Scores

- We need a way to aggregate scores across individuals so that we can draw some general conclusions. We define a *cumulative score process*, which is advantageous because we know its distribution under H_0:
  - Order individuals by the auxiliary variable.
  - Define $t \in (0, 1)$. The empirical cumulative score process is defined by:

$$\mathbf{B}(t; \hat{\theta}) = \widehat{\mathbf{I}}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} s(\hat{\theta}; \mathbf{x}_i).$$

    where $\lfloor nt \rfloor$ is the integer part of nt and $\widehat{\mathbf{I}}$ is some consistent covariance matrix estimate, e.g., the observed information matrix $\mathbf{I}(\widehat{\theta})$.

Tests

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

$$\mathbf{B}(\cdot; \widehat{\boldsymbol{\theta}}) \xrightarrow{d} \mathbf{B}^0(\cdot),$$

  where $\mathbf{B}^0(\cdot)$ 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.

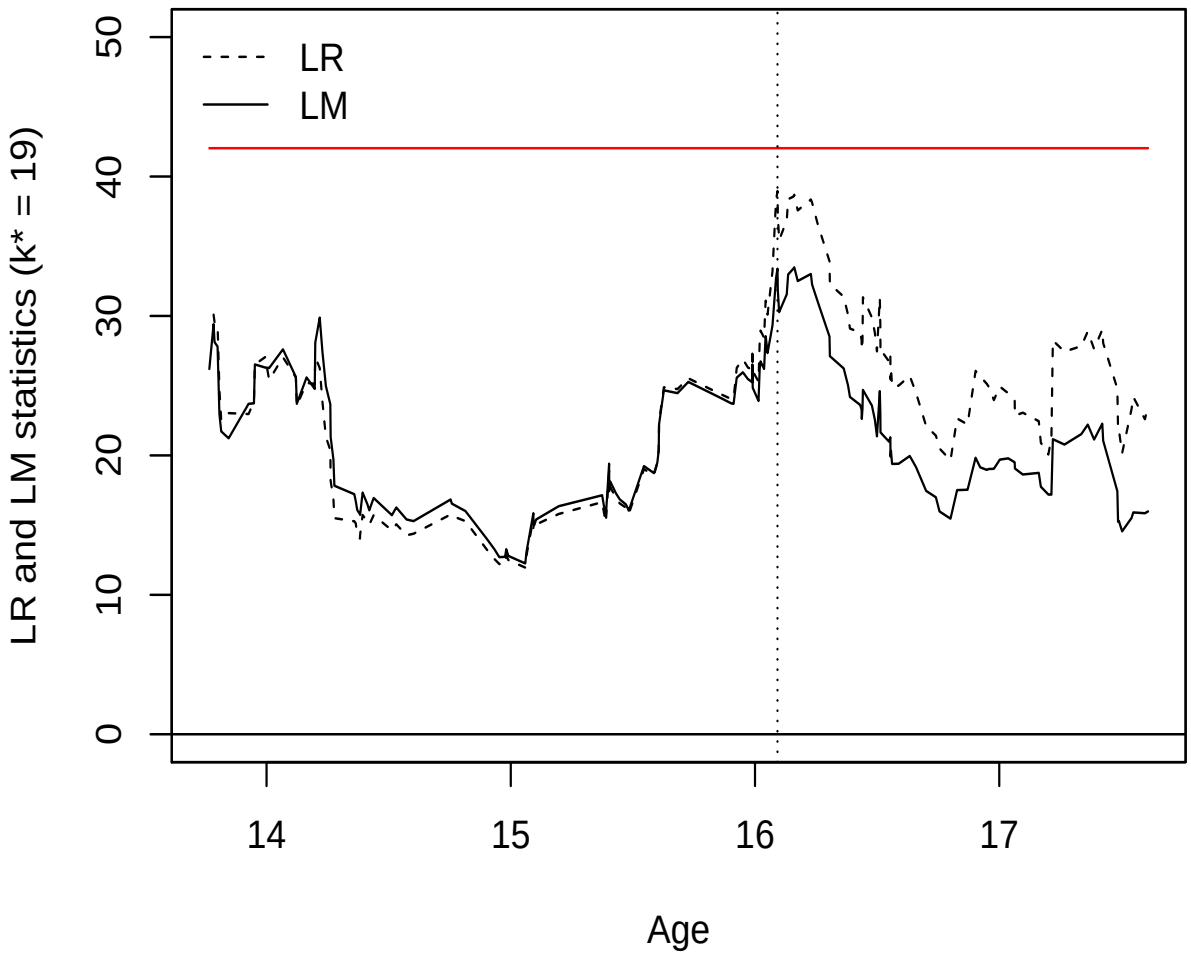
Tests

- Test statistics aggregate over parameters (columns) $j = 1, \ldots, k$ and observations (rows) $i = 1, \ldots, n$, employing different norms.
- Special cases:

$$
\begin{aligned}
DM &= \max_{i=1,\ldots,n} \max_{j=1,\ldots,k} |\mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}|, \\
CvM &= n^{-1} \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,k} \mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \\
\max LM &= \max_{i=\underline{i},\ldots,\bar{i}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\ldots,k} \mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}^2.
\end{aligned}
$$

- $\mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}$ is short for $\mathbf{B}(i/n; \hat{\boldsymbol{\theta}})_j$.

Illustration



Ordinal vs Continuous Tests

- Last year, we showed that the above tests have adequate power to be useful in practice.
- However, the tests require a unique ordering of individuals w.r.t. the auxiliary variable.
- If the auxiliary variable is ordinal, we have a large number of ties. What to do?

Tests along Ordinal Variables

- To obtain a test statistic in the ordinal case, we allow all individuals with the same value of the auxiliary variable to simultaneously enter into the cumulative sum. We then apply the same functionals that were used in the continuous case.
- Critical values are obtained by summing bins of a Brownian bridge, where bin sizes match the observed bin sizes of the ordinal variable. This allows us to treat ordinal auxiliary variables as truly ordinal, which no existing methods (to our knowledge) can do.

Tests along Ordinal Variables

- Instead of aggregating over all $i = 1, \ldots, n$, aggregate only over $i_\ell = \lfloor n \cdot t_\ell \rfloor$ for the cumulative proportions t_ℓ $(\ell = 1, \ldots, m)$.
- Tests: Obtain ordinal versions of the max LM test and (weighted) double maximum test.

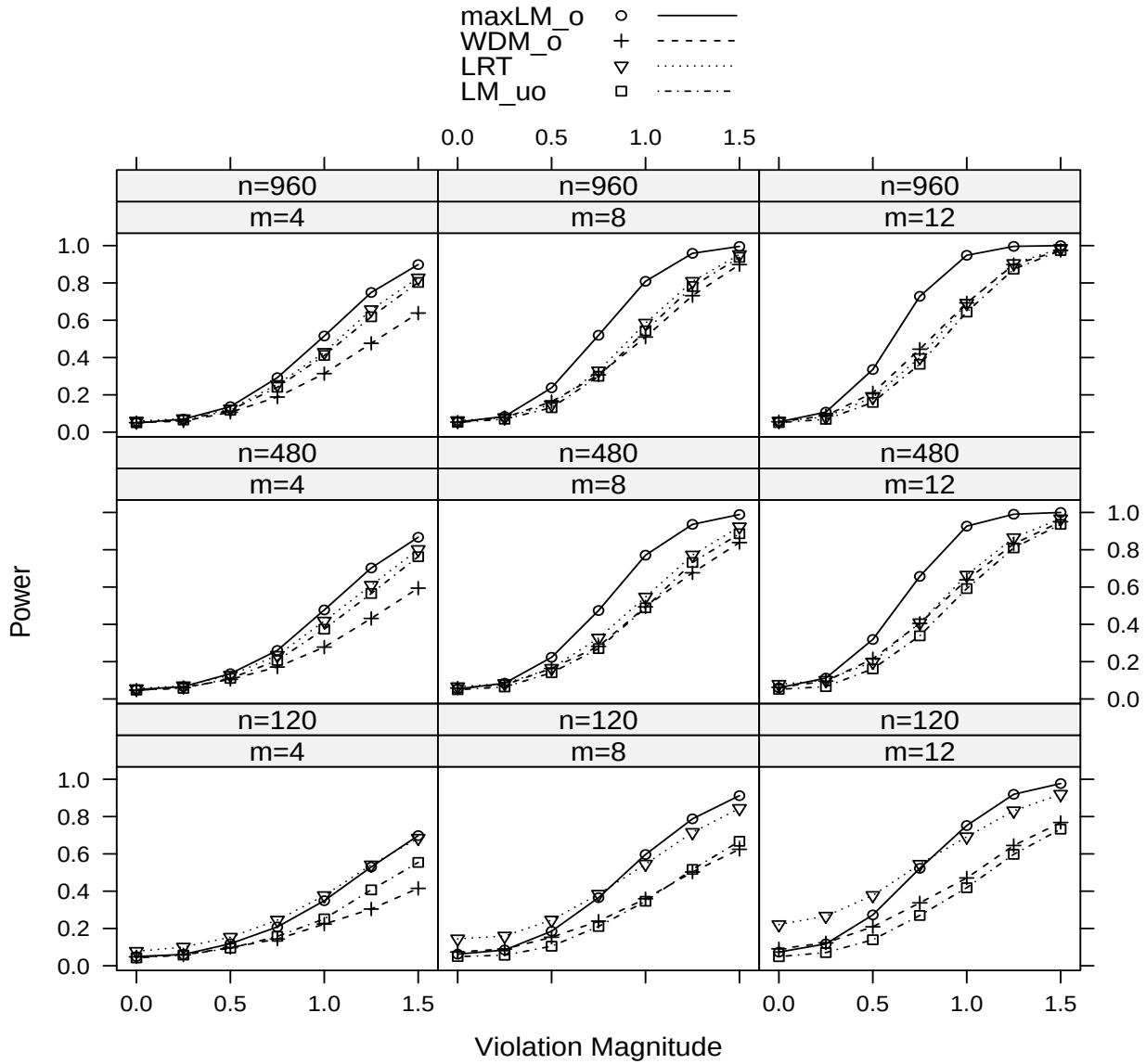
$$WDM_o = \max_{i \in \{i_1, \ldots, i_m\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \max_{j=1,\ldots,k} |\mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}|,$$

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

Simulation 1

- Simulation 1:
  - Two-factor model, with each factor having three unique indicators.
  - Measurement invariance violation in three unique variance parameters, growing as we move up the ordinal variable.
  - Sample size in $\{120, 480, 960\}$.
  - Levels of the ordinal variable in $\{4, 8, 12\}$.
  - Four test statistics (two new statistics, the multiple group LRT, and the unordered Lagrange multiplier test).

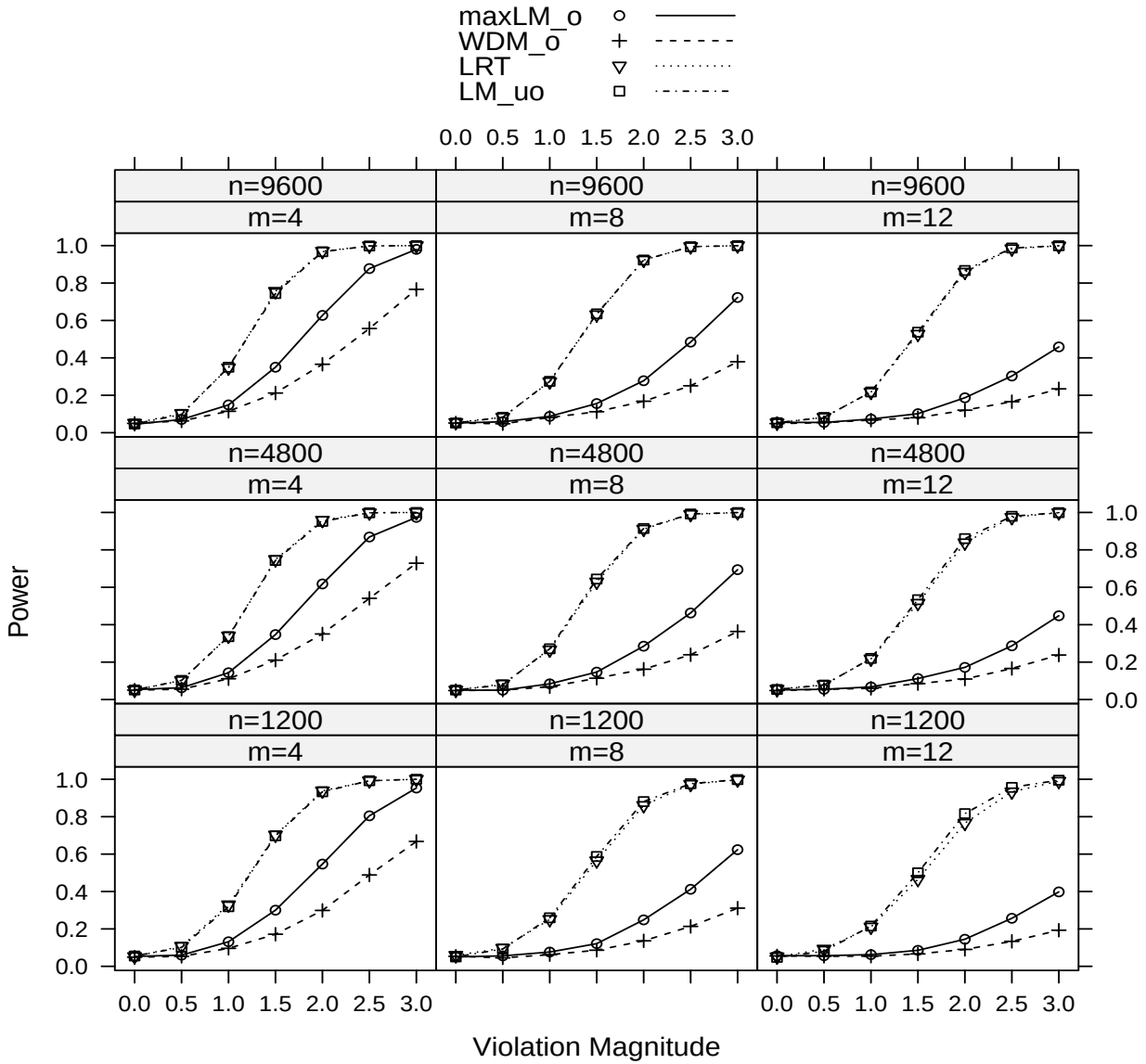
Simulation 1



Simulation 2

- Simulation 2: Are the proposed test statistics less sensitive to large n, as compared to the analogous likelihood ratio test (e.g., Bentler and Bonnett, 1980)?
  - Same two-factor model from before.
  - Small measurement invariance violation (0, 0.5, . . . , 3 standard errors apart) in all unique variance parameters at a middle level of the ordinal auxiliary variable.
  - Sample size in $\{1200, 4800, 9600\}$.
  - Levels of the ordinal variable in $\{4, 8, 12\}$.
  - Four test statistics (two new statistics, the multiple group LRT, and the unordered Lagrange multiplier test).

Simulation 2



Simulations

- Conclusions from ordinal simulations
  - In the ordinal scenarios examined, the ordered Lagrange multiplier statistic has higher power to detect measurement invariance violations than does the traditional LRT.
  - The ordered Lagrange multiplier statistic is also less sensitive to minor invariance violations at large n, as compared to the traditional LRT (the ordered double-max statistic is also less sensitive, but its power is generally poor).
  - The ordered Lagrange multiplier statistic has good potential for practical application (though a downside is that critical values must be simulated).

Example

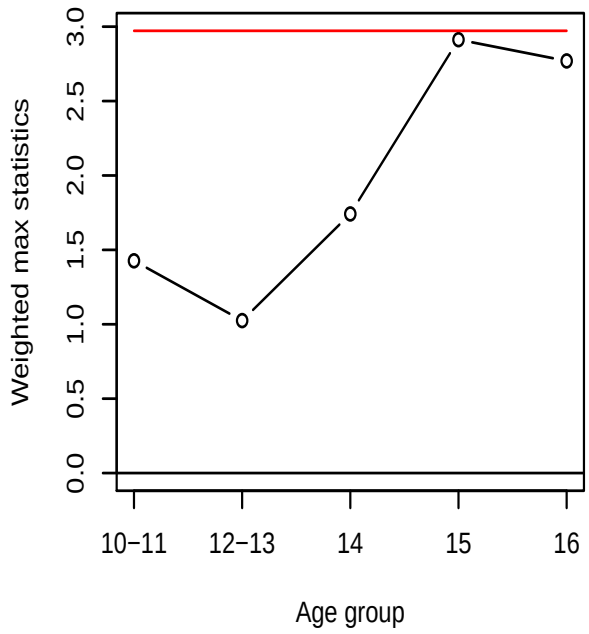
- Example: A small number of scales have been developed to measure gratitude in adults. Can these scales also be used with youth? (Froh, Fan, et al., 2011)
  1. $n = 1,401$ youth ages 10–19 complete three gratitude scales. Age groups are {10–11, 12–13, 14, 15, 16, 17–19}.
  2. Measurement invariance examined via factor analyses of each individual scale (one-factor models), with increasingly-restricted parameters.

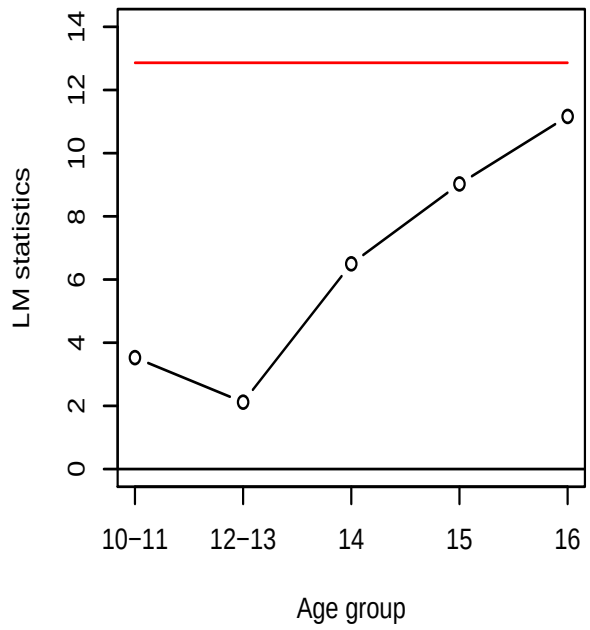
Example

- The large sample size made the authors suspicious about significant likelihood ratio tests of model fit, making it difficult to draw final conclusions about measurement invariance.
  - GQ6 scale: Compare congeneric model to tau-equivalent model, obtain $\chi^2_{20} = 38.08, p = .009$
  - GAC scale: Compare tau-equivalent model to parallel model, obtain $\chi^2_{20} = 167.72, p < .001$
- To use our proposed statistics in each case, we fit the restricted model and test the parameters that are freed in the less-restricted model.

GQ6 Results

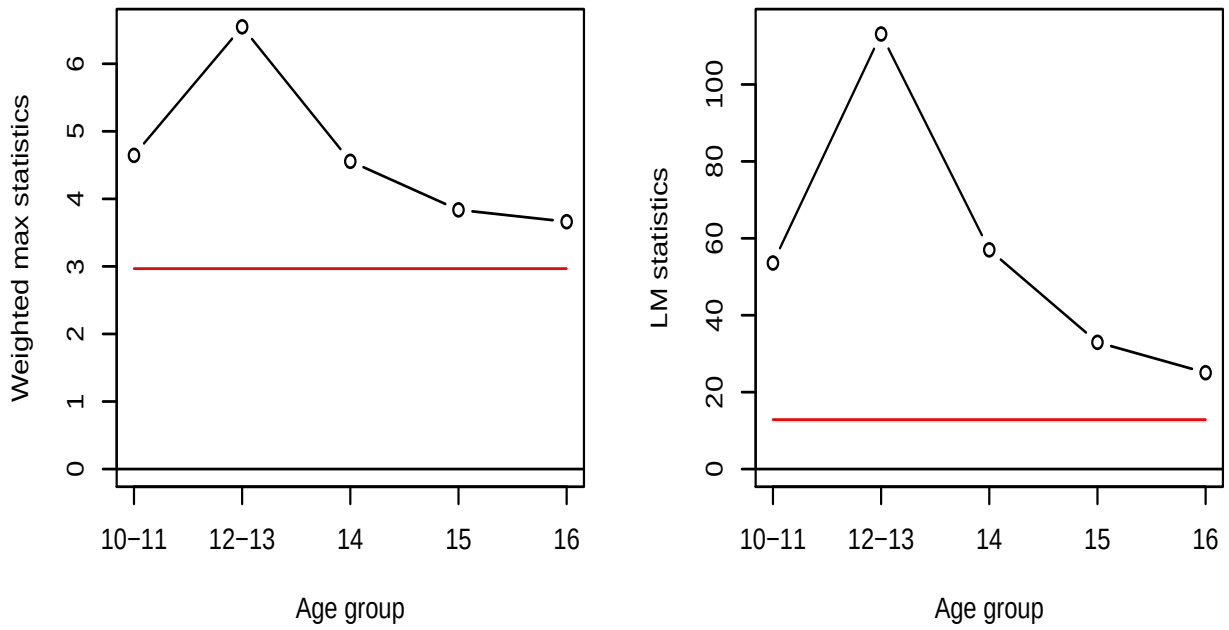
- LRT implies the congeneric model is better, whereas our proposed statistics imply the tau-equivalent model is as good ($WDM_o = 2.91, p = .06$; $\max LM_o = 11.16, p = .10$)





GAC Results

- LRT implies the tau-equivalent model is better, and our proposed statistics agree ($WDM_o = 6.55, p < .01$; $\max LM_o = 113.13, p < .01$)



General Conclusions

- The family of measurement invariance tests considered here usefully extends to the ordinal case:
  - Increased sensitivity to "ordinal" violations.
  - Decreased sensitivity to anomalous violations induced by large n.
  - Ability to test at low n, where multiple-group models are not feasible (the proposed tests only require the "reduced" model).

Software

- R currently has functionality to carry out the proposed tests for general SEMs:
  - `lavaan` for model estimation, with score extraction included in recent versions via `estfun()`.
  - `strucchange` for carrying out the proposed tests.
  - See `http://semtools.r-forge.r-project.org/` for papers and relevant code.

Acknowledgements

- Support from NSF grant SES-1061334
- Yves Rosseel, `lavaan` development

- Questions?