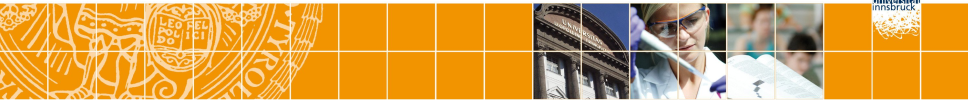
Universität Innsbruck



Score-Based Tests of Measurement Invariance with Respect to Continuous and Ordinal Variables

Achim Zeileis, Edgar C. Merkle, Ting Wang

http://eeecon.uibk.ac.at/~zeileis/

Overview

- Motivation
- Framework
- Score-based tests
    - Continuous variables
    - Ordinal variables
    - Categorial variables
- Software
- Illustrations

Motivation

Psychometric models: Typically measure latent scales based on certain manifest variables, e.g., item response theory (IRT) models or confirmatory factor analysis (CFA).

Crucial assumption: Measurement invariance (MI). Otherwise observed differences in scales cannot be reliably attributed to the latent variable that the model purports to measure.

Parameter stability: In parametric models, the MI assumption corresponds to stability of parameters across all possible subgroups.

Inference: The typical approach for assessing MI is

- to split the data into reference and focal groups,
- assess the stability of selected parameters (all or only a subset) across these groups
- by means of standard tests: likelihood ratio (LR), Wald, or Lagrange multiplier (LM or score) tests.

Motivation

Problems:

- Subgroups have to be formed in advance.
- Continuous variables are often categorized into groups in an ad hoc way (e.g., splitting at the median).
- In ordinal variables the ordering of the categories is often not exploited – assessing only if at least one group differs from the others.
- When likelihood ratio or Wald tests are employed, the model has to be fitted to each subgroup which can become numerically challenging and computationally intensive.

Motivation

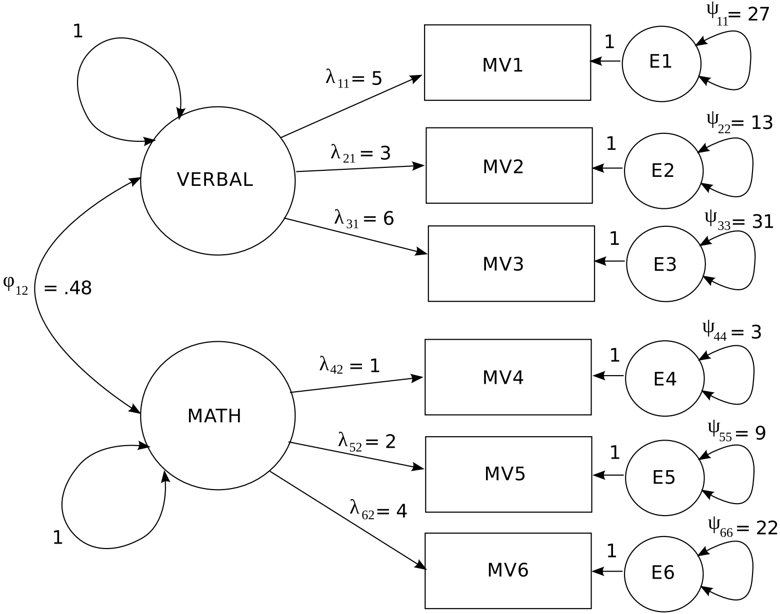
Idea:

- Generalize the LM test.
- Thus, the model only has to be fitted once under the MI assumption to the full data set.
- Catpure model deviations along a variable that is suspected to cause MI violations.
- Exploit ordering to assess if there is (at least) one split so that the model parameters before and after the split differ.
- The split does *not* have to be known or guessed in advance.

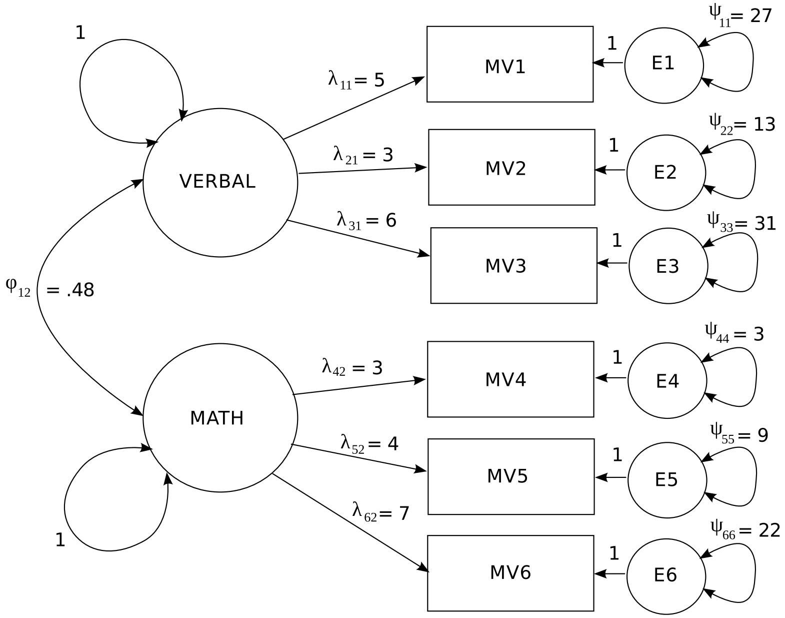
Illustration: CFA for artificial data.

- Model with two latent scales (verbal and math).
- Three manifest variables for each scale.
- Violation of MI for the math loadings along the age of the subjects.

Motivation: CFA for age ≤ 16



Motivation: CFA for age > 16



Framework

Model: Based on log-likelihood $\ell(\cdot)$ for p-dimensional observations $\boldsymbol{x}_i$ ($i = 1, \dots, n$) and k-dimensional parameter $\boldsymbol{\theta}$.

Estimation: Maximum likelihood.

$$\hat{\boldsymbol{\theta}} \; = \; \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i).$$

Equivalently: Solve first order conditions

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

where the score function is the partial derivative of the casewise likelihood contributions w.r.t. the parameters $\boldsymbol{\theta}$.

$$\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{x}_i) \; = \; \left(\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{x}_i)}{\partial \theta_1}, \dots, \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{x}_i)}{\partial \theta_k} \right)^\top.$$

Framework

Assumption: Distribution/likelihood of $\boldsymbol{x}_i$ depends only on the latent scales (through the parameters $\boldsymbol{\theta}$) – but not on any other variable v_i.

Alternative view: Parameters $\boldsymbol{\theta}$ do not depend any such variable v_i. Hence assess for $i = 1, \dots, n$

$$\begin{aligned} H_0 : \boldsymbol{\theta}_i &= \boldsymbol{\theta}_0, \\ H_1 : \boldsymbol{\theta}_i &= \boldsymbol{\theta}(v_i). \end{aligned}$$

Special case: Two subgroups resulting from one split point ν.

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

Tests: LR/Wald/LM tests can be easily employed if pattern $\boldsymbol{\theta}(v_i)$ is known, specifically for H_1^* with fixed split point ν.

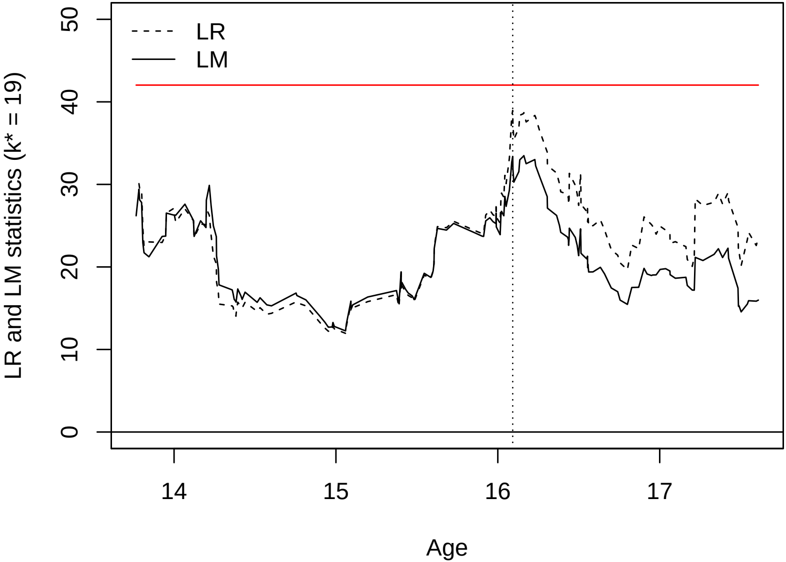
Framework

For unknown split points: Compute LR/Wald/LM tests for each possible split point $v_1 \leq v_2 \leq \dots \leq v_n$ and reject if the maximum statistic is large.

Caution: By maximally selecting the test statistic different critical values are required (not from a χ^2 distribution)!

Illustration: Assess all $k^* = 19$ model parameters from the artificial CFA example along the continuous variable age (v_i).

Framework



Framework

Note: For the maxLM test the parameters $\hat{\theta}$ only have to be estimated once. Only the model scores $\boldsymbol{s}(\hat{\theta}; \boldsymbol{x}_i)$ have to be aggregated differently for each split point.

More generally: Consider a class of tests that assesses whether the model "deviations" $\boldsymbol{s}(\hat{\theta}; \boldsymbol{x}_i)$ depend on v_i. This can consider only a subset k^* of all k parameters/scores or try to capture other patterns than H_1^*.

Score-based tests

Fluctuation process: Capture fluctuations in the cumulative sum of the scores ordered by the variable v.

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

- $\hat{\boldsymbol{I}}$ – estimate of the information matrix.
- t – proportion of data ordered by v.
- $\lfloor n \cdot t \rfloor$ – integer part of $n \cdot t$.
- $x_{(i)}$ – observation with the i-th smallest value of the variable v.

Functional central limit theorem: Under H_0 convergence to a (continuous) Brownian bridge process $\boldsymbol{B}(\cdot; \hat{\theta}) \stackrel{d}{\rightarrow} \boldsymbol{B}^0(\cdot)$, from which critical values can be obtained – either analytically or by simulation.

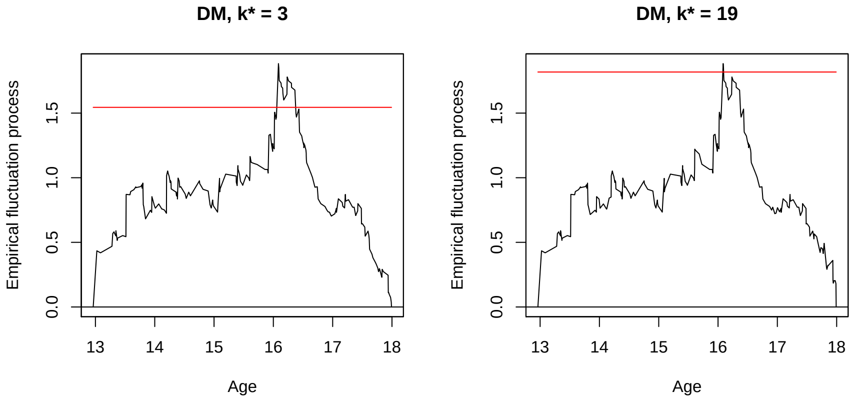
Score-based tests: Continuous variables

Test statistics: The empirical process can be viewed as a matrix $\boldsymbol{B}(\hat{\theta})_{ij}$ with rows $i = 1, \dots, n$ (observations) and columns $j = 1, \dots, k$ (parameters). This can be aggregated to scalar test statistics along continuous the variable v.

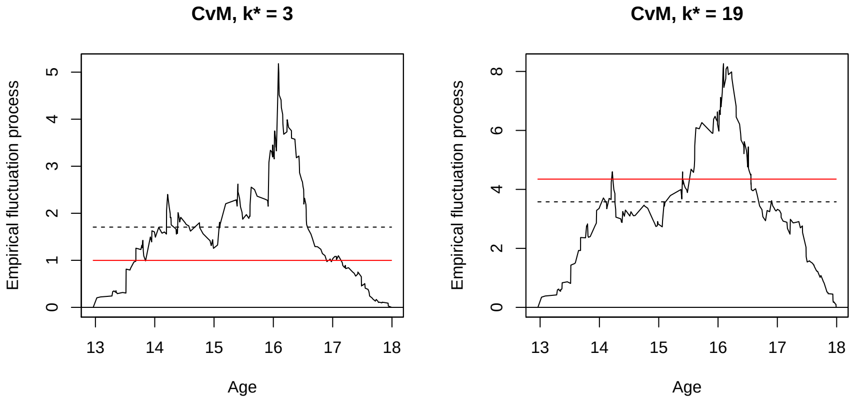
$$
\begin{aligned}
DM &= \max_{i=1,\dots,n} \max_{j=1,\dots,k} |\boldsymbol{B}(\hat{\theta})_{ij}| \\
CvM &= n^{-1} \sum_{i=1,\dots,n} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\theta})_{ij}^2, \\
\max LM &= \max_{i=\underline{i},\dots,\overline{\imath}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right)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

Critical values: Analytically for DM. Otherwise by direct simulation or further refined simulation techniques.

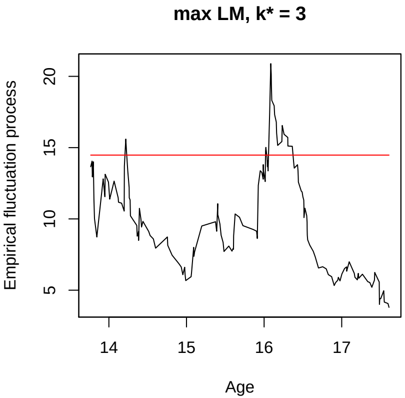
Score-based tests: Continuous variables

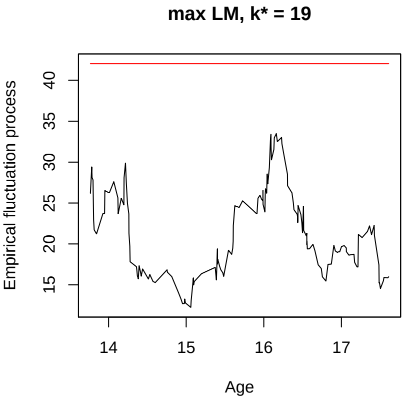


Score-based tests: Continuous variables



Score-based tests: Continuous variables





Score-based tests: Ordinal variables

Test statistics: Aggregation along ordinal variables v with m levels.

$$
\begin{aligned}
WDM_o &= \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1/2} \max_{j = 1, \dots, k} |\boldsymbol{B}(\hat{\theta})_{ij}|, \\
\max LM_o &= \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j = 1, \dots, k} \boldsymbol{B}(\hat{\theta})_{ij}^2,
\end{aligned}
$$

where $i_1, \dots, i_{m-1}$ are the numbers of observations in each category.

Critical values: For WDM_o directly from a multivariate normal distribution. For $\max LM_o$ via simulation.

Score-based tests: Categorical variables

Test statistic: Aggregation within the m (unordered) categories of v.

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

Critical values: From a χ^2 distribution (as usual).

Asymptotically equivalent: LR test.

Software

R packages:

- *strucchange* implements this general framwork for parameter instability tests.
- Object-oriented implementation that can be applied to many model classes, including *lavaan* objects for CFA models.
- Other psychometric models that cooperate with *strucchange* are provided in *psychotools*, e.g., IRT models (Rasch, partial credit, rating scale), Bradley-Terry models for paired comparisons, and multinomial processing tree models.
- Model-based recursive partitioning based on the general parameter instability tests are provided in *partykit*.
- Adaptation to psychometric models in *psychotree*.

CFA: Youth gratitude

Question: Does measurement invariance hold across age groups when an adult gratitude scale is applied to youth subjects?

Source: Froh *et al.* (2011, *Psychological Assessment*). "Measuring Gratitude in Youth: Assessing the Psychometric Properties of Adult Gratitude Scales in Children and Adolescents."

Data:

- GQ-6 gratitude scale with five Likert scale items of seven points each.
- Application to $n = 1401$ youth aged 10–19 years (six age groups).
- Assess the factor loadings of a one-factor model.

CFA: Youth gratitude

Packages:

```
R> library("lavaan")
R> library("strucchange")
```

Data: Omitting incomplete cases.

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

Estimation: One-factor CFA with loadings restricted to be equal across age groups.

```
R> gq6_cfa <- cfa("f1 =~ gq6_1 + gq6_2 + gq6_3 + gq6_4 + gq6_5",
+    data = yg, group = "agegroup", meanstructure = TRUE,
+    group.equal = "loadings")
```

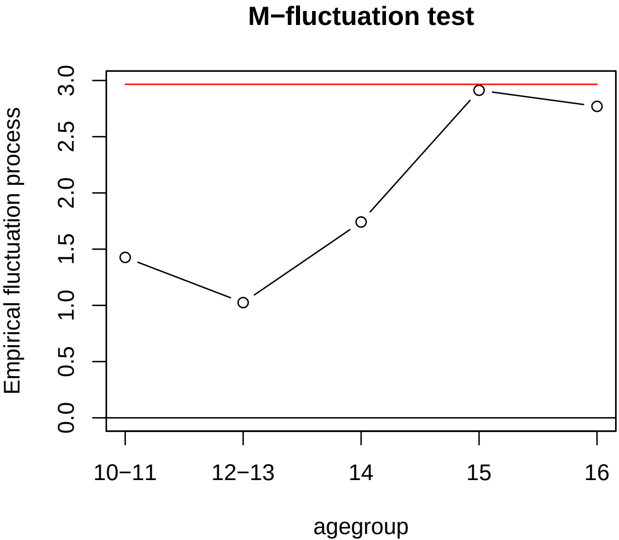
CFA: Youth gratitude

Measurement invariance tests:

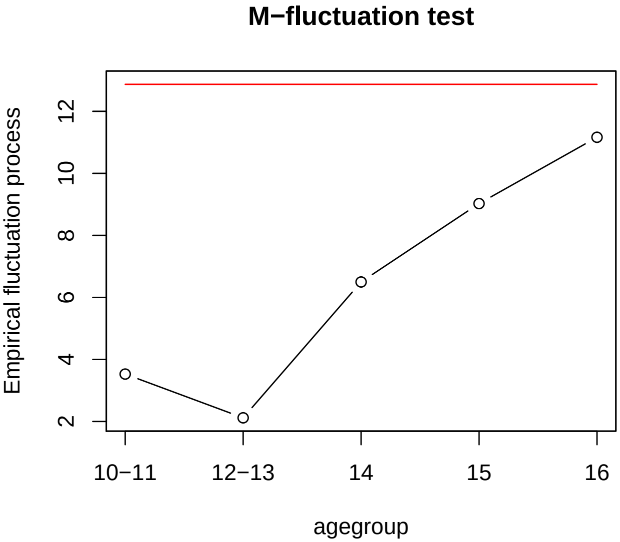
```
R> sctest(gq6_cfa, order.by = yg$agegroup, parm = 1:4,
+    vcov = "info", functional = "WDMo", plot = TRUE)

M-fluctuation test
data:  gq6_cfa
f(efp) = 2.9129, p-value = 0.0591

R> sctest(gq6_cfa, order.by = yg$agegroup, parm = 1:4,
+    vcov = "info", functional = "maxLMo", plot = TRUE)

M-fluctuation test
data:  gq6_cfa
f(efp) = 11.163, p-value = 0.09624
```

Both tests reflect only moderate parameter instability across age groups and do not show significant violations of measurement invariance at 5% level.

CFA: Youth gratitude



CFA: Youth gratitude



IRT: Examining exams

Question: Does measurement invariance hold in a Rasch model for single-choice exam results?

Source: Mathematics for first-year business and economics students at Universität Innsbruck. Online tests (conducted in OpenOLAT) and written exams for 500–1,000 students per semester.

Data: Individual results from an end-term exam.

- 729 students (out of 941 registered).
- 13 single-choice items with five answer alternatives, covering the basics of analysis, linear algebra, financial mathematics.
- Two groups with partially different item pools (on the same topics). Individual versions of items generated via *exams*.
- Correctly solved items yield 100% of associated points. Items without correct solution can either be unanswered (0%) or with an incorrect answer (-25%). Only considered as binary here.

IRT: Examining exams

Packages:

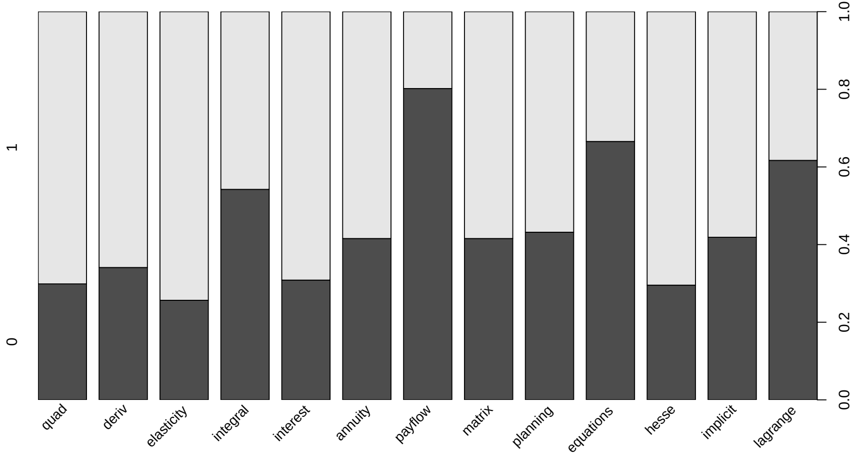
```
R> library("psychotools")
R> library("psychotree")
```

Data: Load, select first group, and exclude extreme scorers.

```
R> load("MathExam.rda")
R> mex <- subset(MathExam, group == 1 & nsolved > 0 & nsolved < 13)
```

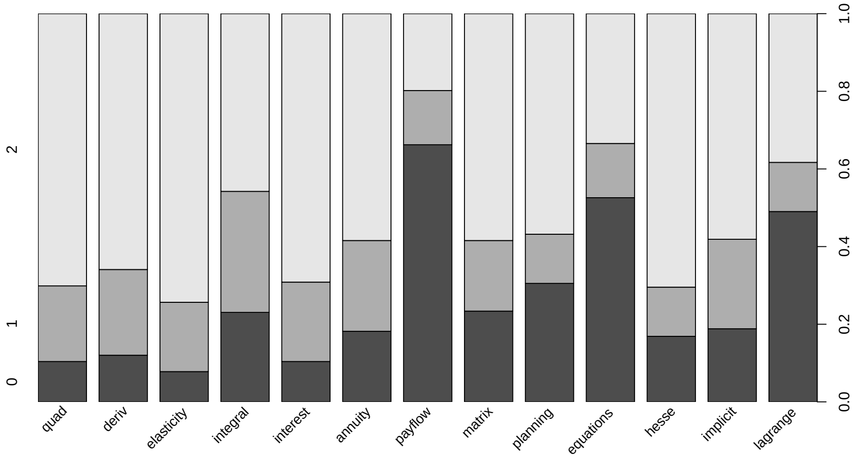
IRT: Examining exams

```
R> plot(mex$solved)
```



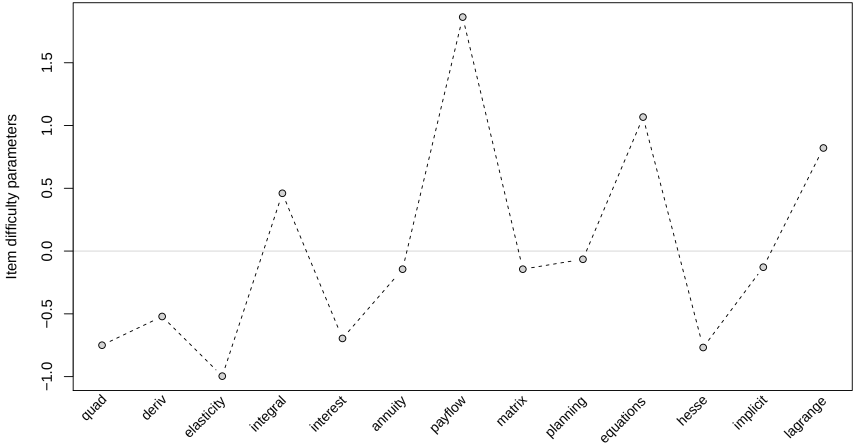
IRT: Examining exams

```
R> plot(mex$credits)
```



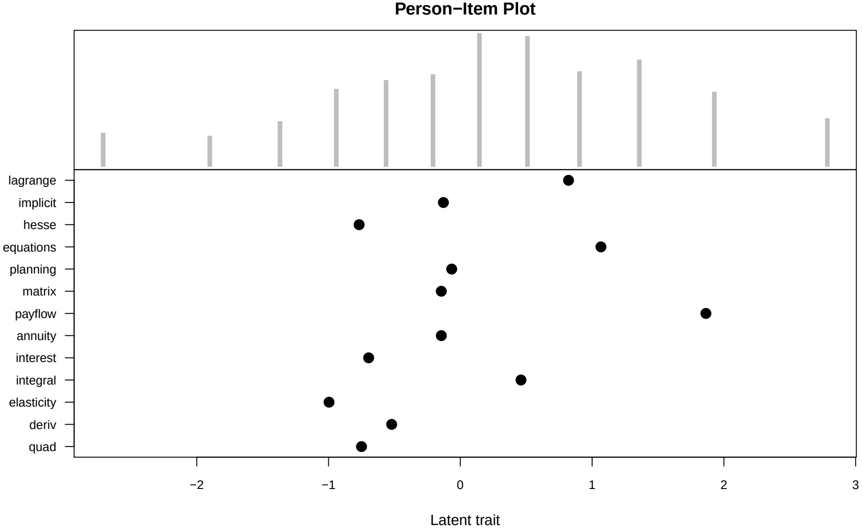
IRT: Examining exams

```
R> mex_rasch <- raschmodel(mex$solved)
R> plot(mex_rasch, type = "profile")
```



IRT: Examining exams

```
R> plot(mex_rasch, type = "piplot")
```



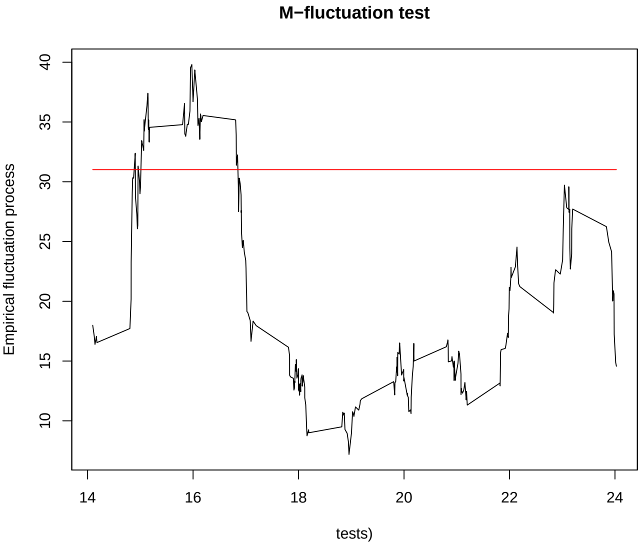
IRT: Examining exams

Measurement invariance tests:

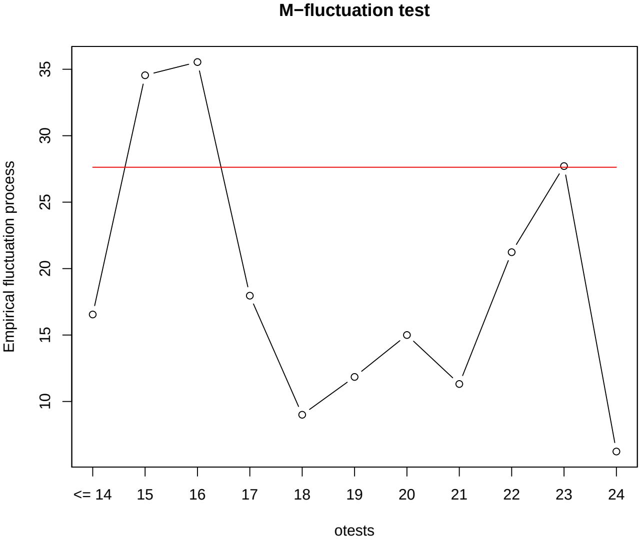
```
R> sctest(mex_rasch, order.by = jitter(mex$tests),
+    vcov = "info", functional = "maxLM", plot = TRUE)

M-fluctuation test
data:  mex_rasch
f(efp) = 39.8, p-value = 0.003047

R> mex$otests <- cut(mex$tests, breaks = c(0, 14:24, 26),
+    ordered = TRUE, labels = c("<= 14", 15:24, ">= 25"))
R> sctest(mex_rasch, order.by = mex$otests,
+    vcov = "info", functional = "maxLMo", plot = TRUE)

M-fluctuation test
data:  mex_rasch
f(efp) = 35.543, p-value = 0.003717
```

Clear violation of measurement invariance: Students that performed poorly in the previous online tests have a different item profile.

IRT: Examining exams

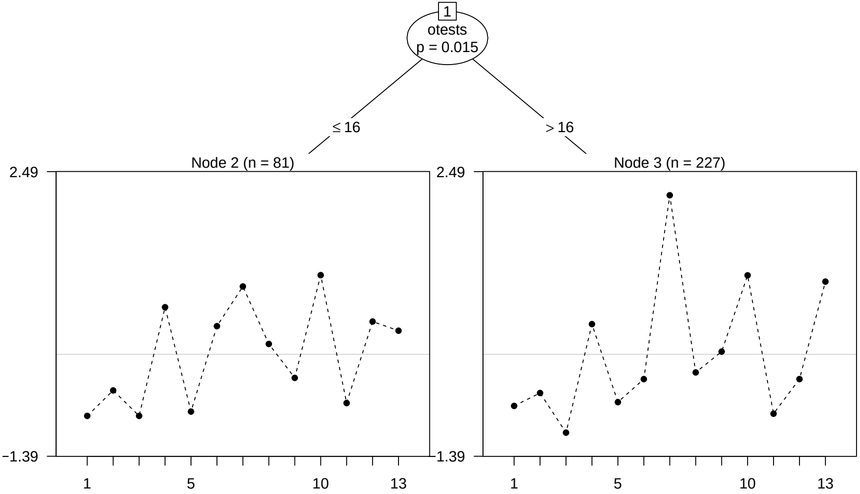


IRT: Examining exams



IRT: Examining exams

```
R> mex_tree <- raschtree(solved ~ otests + attempt + semester + study,
+     data = mex, vcov = "info", ordinal = "L2")
```



Paired comparisons: Modeling topmodels

Question: Does measurement invariance hold for a Bradley-Terry preference scaling of attractiveness?

Source: Strobl, Wickelmaier, Zeileis (2010, *Journal of Educational and Behavioral Statistics*). "Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning."

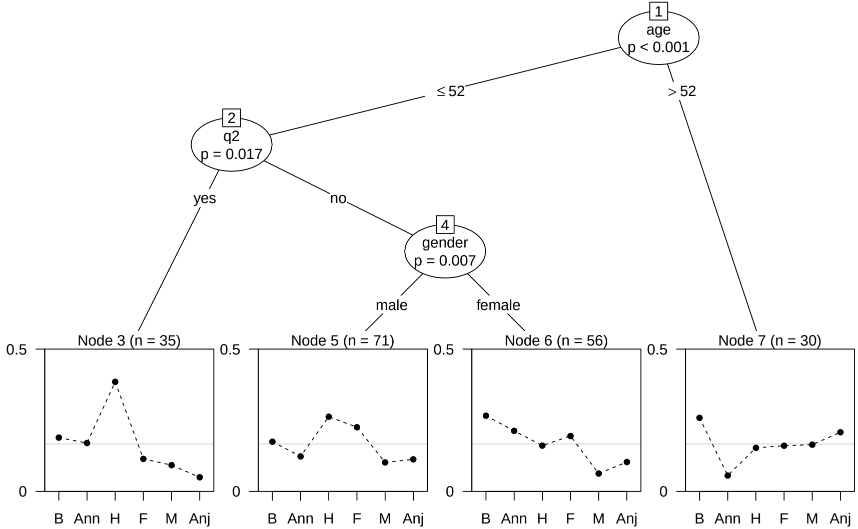
Data:

- Paired comparisons of attractiveness for *Germany's Next Topmodel 2007* finalists: Barbara, Anni, Hana, Fiona, Mandy, Anja.
- Survey with 192 respondents at Universität Tübingen.
- Available covariates: Gender, age, familiarty with the TV show.
- Familiarity assessed by yes/no questions: (1) Do you recognize the women?/Do you know the show? (2) Did you watch it regularly? (3) Did you watch the final show?/Do you know who won?

Paired comparisons: Modeling topmodels

Paired comparisons: Modeling topmodels

```
R> data("Topmodel2007", package = "psychotree")
R> tm_mob <- bttree(preference ~ ., data = Topmodel2007,
+     minsplit = 5, ref = "Barbara")
```



Paired comparisons: Modeling topmodels

Recursively partitioned preferences: Standardized ranking from Bradley-Terry model.

	Barbara	Anni	Hana	Fiona	Mandy	Anja
3	0.19	0.17	0.39	0.11	0.09	0.05
5	0.17	0.12	0.26	0.23	0.10	0.11
6	0.27	0.21	0.16	0.19	0.06	0.10
7	0.26	0.06	0.15	0.16	0.16	0.21

Summary

- General score-based test framework for assessing measurement invariance in parametric psychometric models.
- Assessment is along some variable *v* which can be continuous, ordinal, or categorical.
- Tests can be seen as generalizations of the Lagrange multiplier test.
- Computation of critical values might require simulation from certain stochastic processes (Brownian bridges).
- Easy-to-use implementation available in R package *strucchange*.
- Can be re-used in model-based recursive partitioning in R packages *partykit* and *psychotree*.

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

References

Merkle EC, Zeileis A (2013). "Tests of Measurement Invariance without Subgroups: A Generalization of Classical Methods." *Psychometrika*, **78**(1), 59–82. doi:10.1007/s11336-012-9302-4

Merkle EC, Fan J, Zeileis A (2014). "Testing for Measurement Invariance with Respect to an Ordinal Variable." *Psychometrika*, **79**(4), 569–584. doi:10.1007/s11336-013-9376-7

Wang T, Merkle EC, Zeileis A (2014). "Score-Based Tests of Measurement Invariance: Use in Practice." *Frontiers in Psychology*, **5**(438). doi:10.3389/fpsyg.2014.00438

Strobl C, Julia Kopf, Zeileis A (2015). "Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model." *Psychometrika*. **80**(2), 289–316. doi:10.1007/s11336-013-9388-3

Strobl C, Wickelmaier F, Zeileis A (2011). "Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning." *Journal of Educational and Behavioral Statistics*, **36**(2), 135–153. doi:10.3102/1076998609359791

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514. doi:10.1198/106186008X319331