Universität Innsbruck

Examining Exams Using Rasch Models and Assessment of Measurement Invariance

Achim Zeileis

`https://eeecon.uibk.ac.at/~zeileis/`

Overview

- Topics
  - Large-scale exams
  - Item response theory with Rasch model
  - Assessment of measurement invariance
- Mathematics 101 exam at Universität Innsbruck
  - Classical tests
  - Anchor methods
  - Score-based tests
  - Model-based recursive partitioning
  - Finite mixture models
- Discussion

Many collaborators: Hannah Frick, Bettina Grün, Kurt Hornik, Torsten Hothorn, Basil Komboz, Julia Kopf, Friedrich Leisch, Edgar C. Merkle, Carolin Strobl, Nikolaus Umlauf, Ting Wang, Florian Wickelmaier.

Large-scale exams

Motivation:

- Statisticians often teach large lecture courses for other fields.
- Statistics, probability, or mathematics in curricula such as business and economics, social sciences, psychology, etc.
- At WU Wien and Universität Innsbruck: Some courses are attended by more than 1,000 students per semester.
- Several lecturers teach lectures and tutorials in parallel.

Typical exams:

- Multiple choice or single choice.
- Evaluated and graded automatically.
- Little further examination of results (if any).

Large-scale exams

Potential questions:

- Ability of students.
- Difficulty of exercises (or items).
- Differential item functioning (DIF).
- Unidimensionality.

At WU: Multiple-choice monitor by Ledermüller, Nettekoven, Weiler/Krakovsky.

Here:

- Rasch model for binary single-choice items.
- Assessment of measurement invariance vs. DIF.

IRT with Rasch model

Motivation: Item response theory (IRT) with Rasch model.

- Measure a single latent trait (here: ability in exam).
- Based on binary items y_{ij} (here: solved correctly vs. not).
- Align person's ability θ_i ($i = 1, \dots, n$) and item's difficulty β_j ($j = 1, \dots, m$) on the same scale.

Model:

$$\begin{aligned} \pi_{ij} &= \text{Pr(person } i \text{ solves item } j) \; = \; \Pr(y_{ij} = 1) \\ \text{logit}(\pi_{ij}) &= \theta_i - \beta_j \end{aligned}$$

- Interval scale with arbitrary zero point.
- Fix reference point by zero constraint (e.g., for β_1 or $\sum_j \beta_j$).
- Consistent estimation via conditional maximum likelihood.
- Sufficient statistics for θ_i: Sum of correct items for person i.

Assessment of measurement invariance

Crucial assumption: Measurement invariance (MI). Otherwise observed differences 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.

Assessment of measurement invariance

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 category ordering 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.

Conceivable solutions:

- Score-based tests “along” numerical/ordinal/categorical covariates.
- Recursive partitioning to capture covariate interactions.
- Finite mixture models without covariates.

Mathematics 101 at Universität Innsbruck

Course: Mathematics for first-year business and economics students at Universität Innsbruck.

Format: Biweekly online tests (conducted in OpenOLAT) and two written exams for about 1,000 students per semester.

Here: 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* in R.
- Correctly solved items yield 100% of associated points. Items without correct solution can either be unanswered (0%) or with an incorrect answer (-25%). Considered as binary here.

Mathematics 101 at Universität Innsbruck

Variables: In `MathExam14W`.

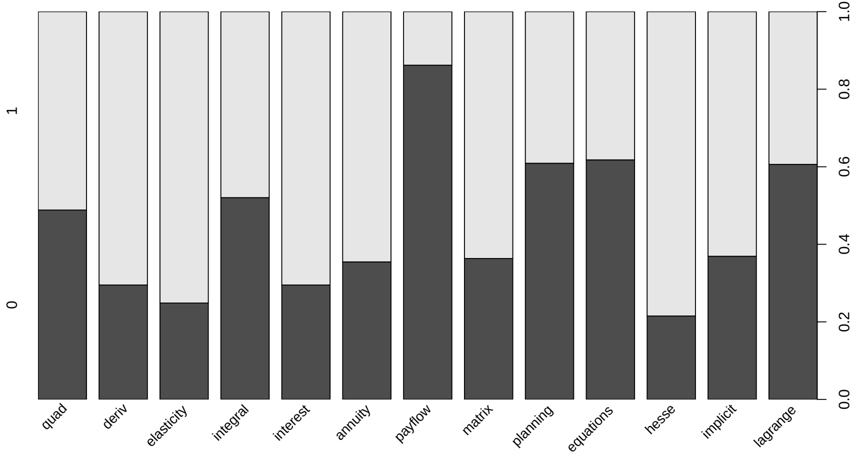
- `solved`: Item response matrix (1/0 coding).
- `group`: Factor for group.
- `tests`: Number of previous online exercises solved (out of 26).
- `nsolved`: Number of exam items solved (out of 13).
- `gender`, `study`, `attempt`, `semester`, ...

In R: Load package/data and exclude extreme scorers.

```
R> library("psychotools")
R> data("MathExam14W", package = "psychotools")
R> mex <- subset(MathExam14W, nsolved > 0 & nsolved < 13)
```

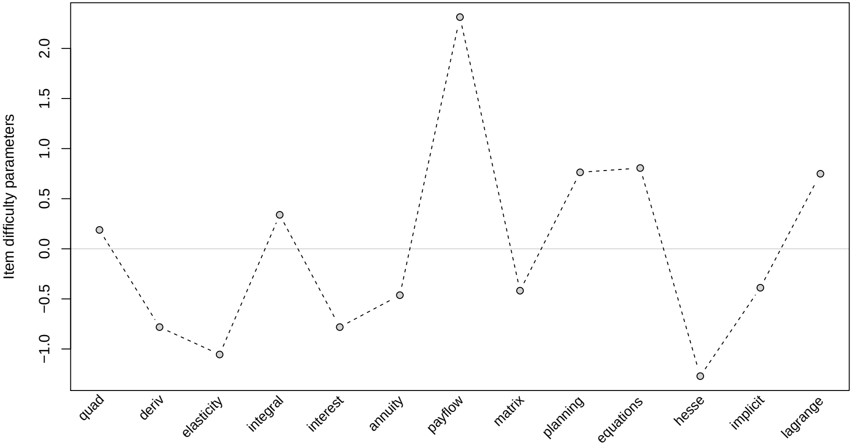
Mathematics 101 at Universität Innsbruck

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



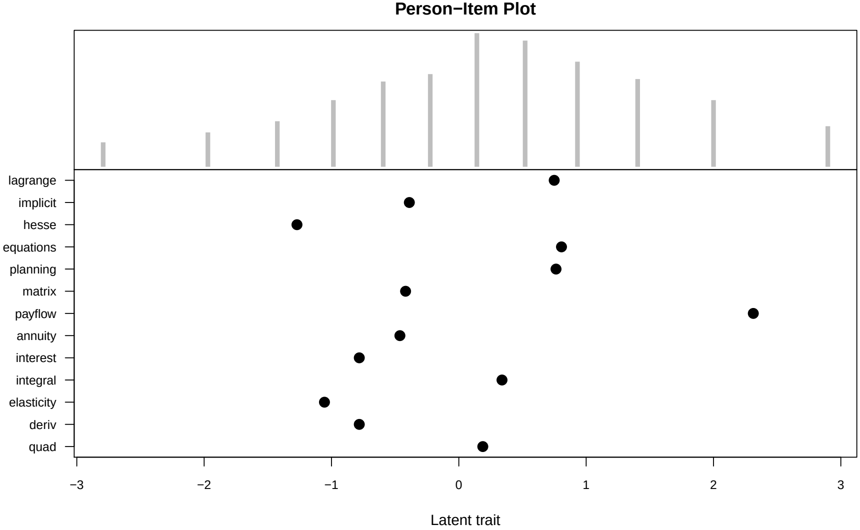
Rasch model

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



Rasch model

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



Classical tests

Of interest: Difference between the two exam groups.

Tests: All χ^2_{12} with 95% critical value 21.0.

- LR: 265.0.
- Wald: 249.4.
- LM/Score: 260.8.

Question: Which items "cause" this DIF?

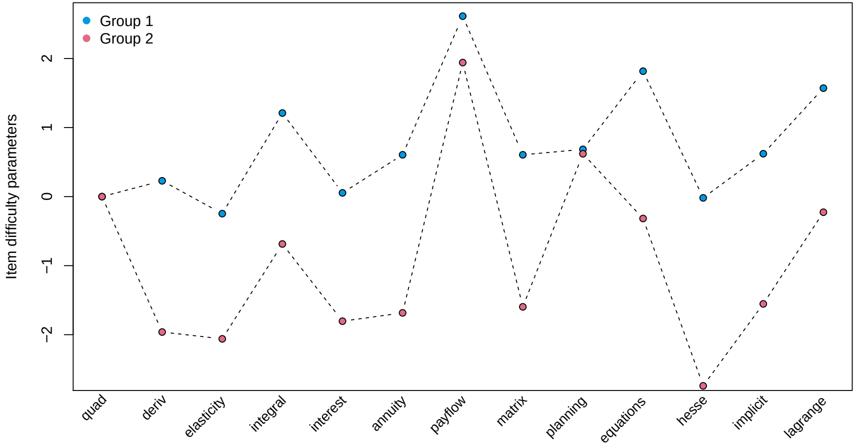
Answer: Use item-wise Wald tests.

$$t_j = \frac{\hat{\beta}_j^{\text{ref}} - \hat{\beta}_j^{\text{foc}}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}^{\text{ref}})_{j,j} + \widehat{\text{Var}}(\hat{\beta}^{\text{foc}})_{j,j}}}.$$

But: "Anchor" items are needed to align the scales from the two groups.

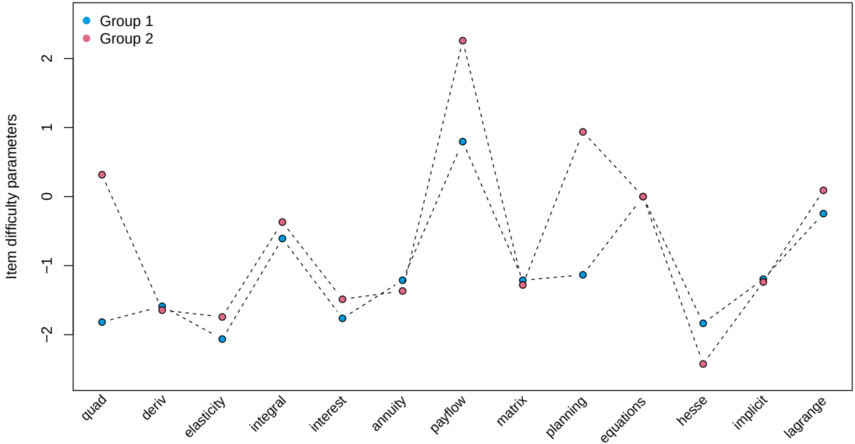
Classical tests

```
R> plot(mr1, parg = list(ref = 1), ...)
R> plot(mr2, parg = list(ref = 1), ...)
```



Classical tests

```
R> plot(mr1, parg = list(ref = 10), ...)
R> plot(mr2, parg = list(ref = 10), ...)
```



Anchor methods

Goal: Select DIF-free anchor items to be able to identify items truly associated with DIF (“chicken or the egg” dilemma).

Approaches: Classes of anchors with different characteristics.

- *All other:* All items – except the item currently studied.
- *Constant:* Predefined number of items (e.g., 1 or 4).
- *Forward:* Iteratively add items.

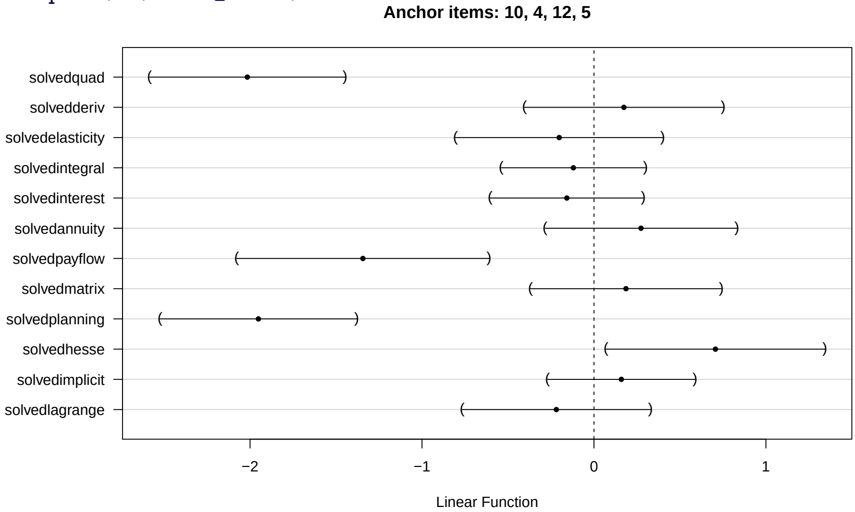
Selection: Rank candidate items based on single-anchor DIF tests.

- Number of significant tests.
- Mean test statistic or *p*-value.
- Mean test statistic or *p*-value beyond median threshold.

Here: Constant anchor class with 4 items and mean *p*-value threshold selection. Single-step adjustment of final inference for multiple testing.

Anchor methods

```
R> ma <- anchortest(solved ~ group, data = mex, adjust = "single-step")
R> plot(ma$final_tests)
```



Score-based tests

Questions:

- Is there further DIF in the two exam groups?
- Is there DIF w.r.t. mathematics ability, e.g., for `tests` $(0, \dots, 13, \dots, 26)$ or `nsolved` $(1, \dots, 12)$?

Problem: Numeric variables without predefined subgroups. Hence, many possible patterns of deviation from parameter stability.

Idea: Generalize the LM test.

- Model only has to be fitted once under the MI assumption to the full data set.
- Catpure model deviations along a variable v that is suspected to cause DIF and violate MI.

Score-based tests

Hypotheses: Under MI parameters $\boldsymbol{\beta}$ do not depend any variable v_i.
Hence assess for $i = 1, \dots, n$

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

Building block: Casewise model deviations.

- Derivative of the casewise log-likelihood w.r.t. the parameters.
- General measure of model deviation (similar to residuals).

$$\boldsymbol{s}(\beta; \boldsymbol{y}_i) = \left(\frac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{y}_i)}{\partial \beta_2}, \dots, \frac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{y}_i)}{\partial \beta_m} \right)^\top$$

Score-based tests

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

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

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

For unknown split point: Compute LR/Wald/LM tests for each possible split point $v_1 \leq v_2 \leq \cdots \leq v_n$ 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)!

More generally: Consider a class of tests that assesses whether the model "deviations" $\boldsymbol{s}(\hat{\beta}; \boldsymbol{y}_i)$ depend on v_i.

Score-based tests

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

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

- $\hat{\boldsymbol{I}}$ – 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 $\boldsymbol{B}(\cdot; \hat{\beta}) \overset{d}{\to} \boldsymbol{B}^0(\cdot)$, 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 $\boldsymbol{B}(\hat{\beta})_{ij}$ with rows $i = 1, \dots, n$ (observations) and columns $j = 1, \dots, m-1$ (parameters). This can be aggregated to scalar test statistics along continuous the variable v.

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

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

Score-based tests: Ordinal variables

Test statistics: Aggregation along ordinal variables v with c categories.

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

where $i_1, \dots, i_{c-1}$ 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 c (unordered) categories of v.

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

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

Asymptotically equivalent: LR test.

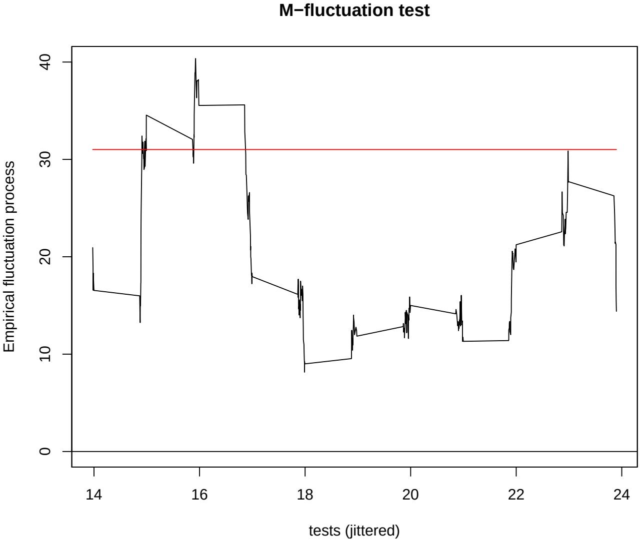
Score-based tests

Here: Test for DIF along `tests` in `group` 1 with max *LM* test (continuous vs. ordinal).

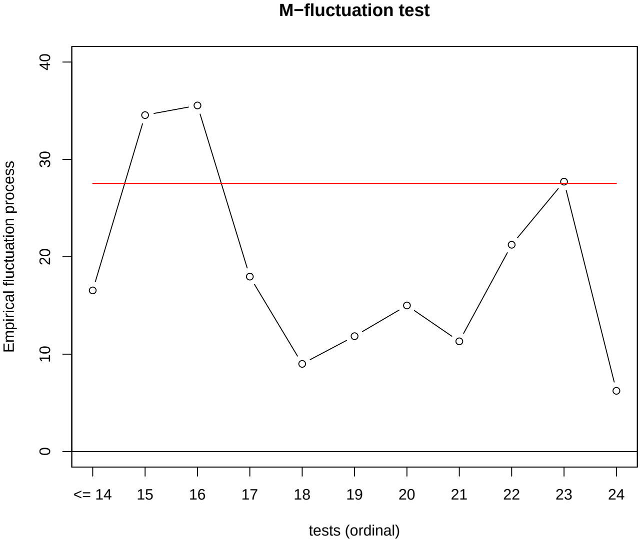
Result: Clear evidence for DIF. Students that performed poorly in the previous online tests have a different item profile.

```
R> library("strucchange")
R> mex1 <- subset(mex, group == 1)

R> sctest(mr1, order.by = mex1$tests, vcov = "info",
+    functional = "maxLM")

        M-fluctuation test

data:  mr1
f(efp) = 40.365, p-value = 0.002508

R> sctest(mr1, order.by = mex1$tests, vcov = "info",
+    functional = "maxLMo")

        M-fluctuation test

data:  mr1
f(efp) = 35.543, p-value = 0.003961
```

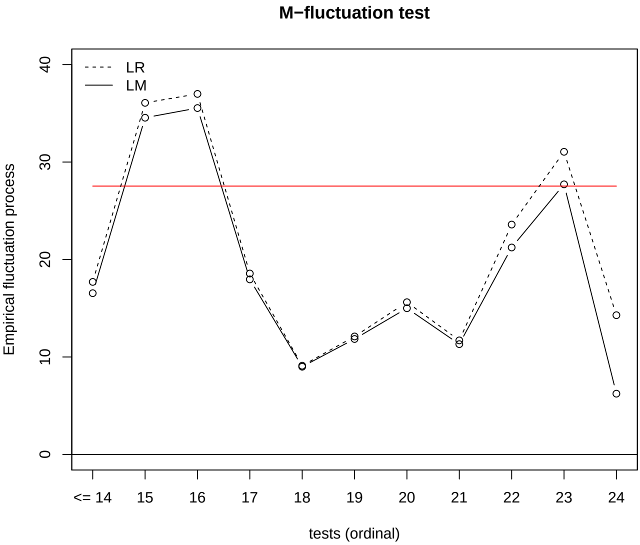
Score-based tests



Score-based tests



Score-based tests



Recursive partitioning

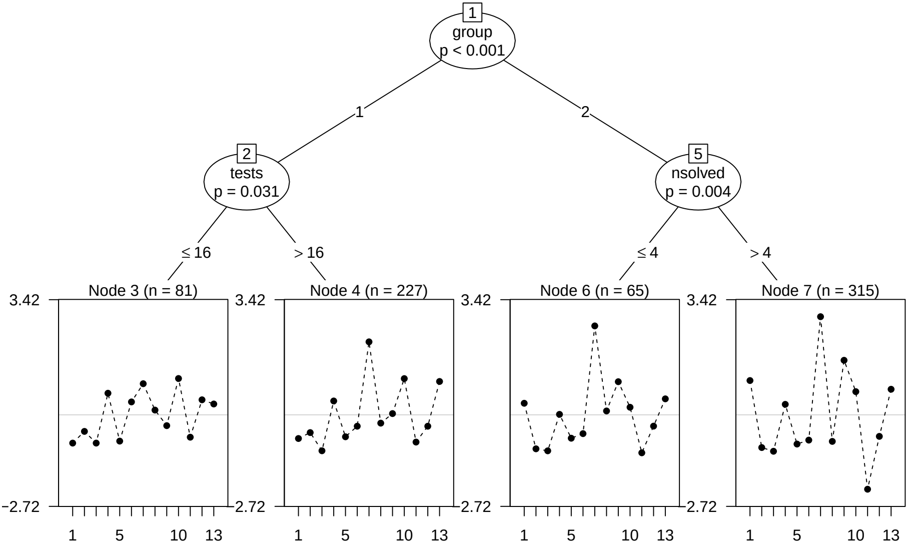
Idea: Apply tests recursively.

- Asess all covariates of interest using Bonferroni adjustment.
- Split w.r.t. covariate with smallest significant *p*-value.
- Select split point by maximizing the log-likelihood.
- Continue until there are no more significant instabilities (or the sample is too small).

Here: Treat numeric variables with few levels as ordinal. Simulate *p*-values for max LM_o test.

```
R> library("psychotree")
R> mex$tests <- ordered(mex$tests)
R> mex$nsolved <- ordered(mex$nsolved)
R> mex$attempt <- ordered(mex$attempt)
R> mex$semester <- ordered(mex$semester)
R> mrt <- raschtree(solved ~ group + tests + nsolved + gender +
+    attempt + study + semester, data = mex,
+    vcov = "info", minsize = 50, ordinal = "L2", nrep = 1e5)
```

Recursive partitioning



Finite mixture models

Question: How to detect DIF without covariate information (e.g., in `group` 1 without `tests`)?

Answer: Finite mixture of Rasch models with $k = 1, \dots, K$ components. Maximize finite mixture likelihood via EM w.r.t. component-specific weights ω_k and item difficulties $\beta^{(k)}$.

$$\max_{\boldsymbol{\omega}, \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(K)}} \prod_{i=1}^{n} \sum_{k=1}^{K} \omega_k f(\boldsymbol{y}_i; \boldsymbol{\beta}^{(k)})$$

Possible extensions:

- Model selection for the number of components K.
- Concomitant variables for the mixture weights $\boldsymbol{\omega}$.
- Component-specific distributions for the raw scores.

Finite mixture models

Here: 2-component mixture with component-specific raw score distribution (mean-variance specification).

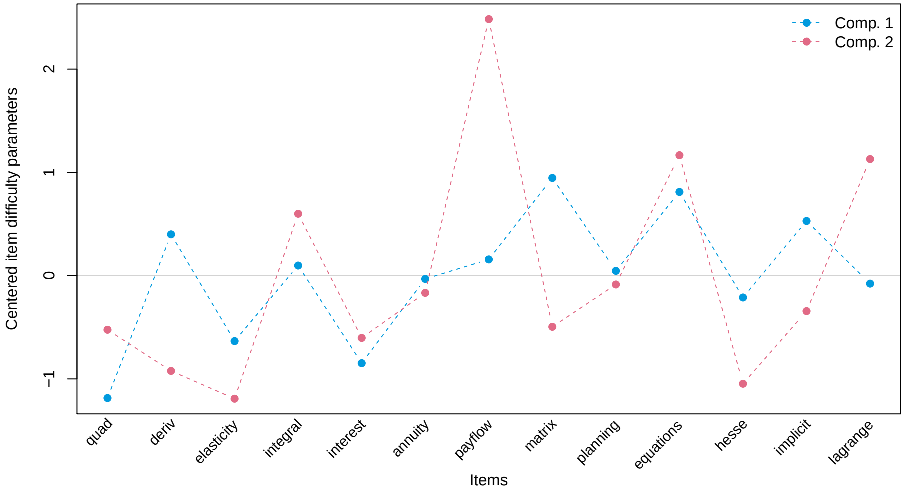
```
R> library("psychomix")
R> mrm <- raschmix(mex1$solved, k = 2, scores = "meanvar")
R> plot(mrm)
```

Result: The “soft” classification found by the mixture model is rather similar to the “hard” split by the tree.

```
R> print(mrm)
```

```
Call:
raschmix(formula = mex1$solved, k = 2, scores = "meanvar")

Cluster sizes:
  1   2
 73 235

convergence after 79 iterations
```

Finite mixture models



Discussion

Summary:

- Flexible toolbox for assessing measurement invariance in parametric psychometric models.
- Detecting violations along one (tests), none (mixture), or many (tree) covariates.
- Exploit different scales of the covariates: continuous, ordinal, or categorical.

Here: Probably quickest overview of DIF patterns with Rasch tree.

At UIBK: Resulting “policy” implications.

- Avoid exam groups if at all possible.
- Seemingly equivalent items can function very differently if students focus their learning on well-known parts of the item pool.

Discussion

R packages:

- *strucchange* provides an object-oriented implementation of the score-based parameter instability tests.
- Model-based recursive partitioning available in *partykit*.
- Psychometric models that cooperate with *strucchange* and *partykit* are provided in *psychotools*: IRT models (Rasch, partial credit, rating scale), Bradley-Terry, multinomial processing trees.
- Psychometric trees in *psychotree*.
- Psychometric mixture models in *psychomix* (based on *flexmix* plus *psychotools*).

Discussion

Exams infrastructure: R package *exams*.

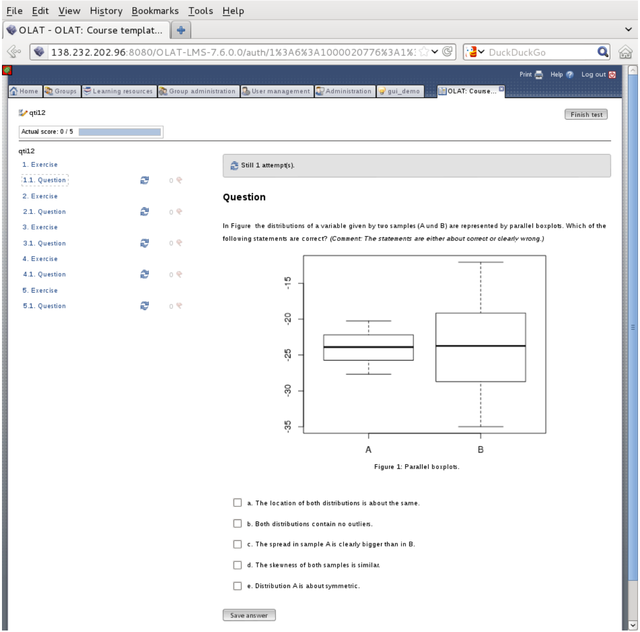
- R for random data generation and computations.
- LaTeX or Markdown for text formatting
- Answer types: Single/multiple choice, numeric, string, cloze.

Output:

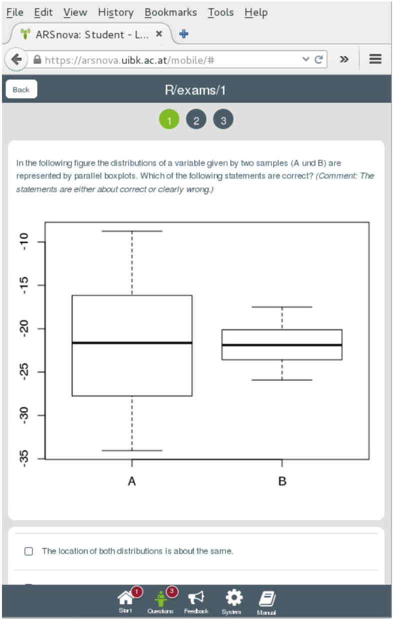
- PDF – either fully customizable or standardized with automatic scanning/evaluation.
- HTML – either fully customizable or embedded into any of the standard formats below.
- Moodle XML.
- QTI XML standard (version 1.2 or 2.1), e.g., for OLAT/OpenOLAT.
- ARSnova, Blackboard, TCExam, WU-Prüfungsserver, . . .

Discussion

Discussion


File Edit View History Bookmarks Tools Help
OLAT - OLAT: Course templat...
138.232.202.96:8080/OLAT-LMS-7.6.0.0/auth/1%3A6%3A1000020776%3A1%3
DuckDuckGo
Print
Help
Log out
Home
Groups
Learning resources
Group administration
User management
Administration
gui_demo
OLAT: Course...
qti12
Finish test
Actual score: 0 / 5
qti12
1. Exercise
1.1. Question
2. Exercise
2.1. Question
3. Exercise
3.1. Question
4. Exercise
4.1. Question
5. Exercise
5.1. Question
Still 1 attempt(s).
Question
In Figure the distributions of a variable given by two samples (A und B) are represented by parallel boxplots. Which of the following statements are correct? (Comment: The statements are either about correct or clearly wrong.)
-15
-20
-25
-30
-35
A
B
Figure 1: Parallel boxplots.
a. The location of both distributions is about the same.
b. Both distributions contain no outliers.
c. The spread in sample A is clearly bigger than in B.
d. The skewness of both samples is similar.
e. Distribution A is about symmetric.
Save answer

Discussion


File Edit View History Bookmarks Tools Help
ARSnova: Student - L...
https://arsnova.uibk.ac.at/mobile/#
Back
R/exams/1
1 2 3
In the following figure the distributions of a variable given by two samples (A und B) are represented by parallel boxplots. Which of the following statements are correct? (Comment: The statements are either about correct or clearly wrong.)
-10
-15
-20
-25
-30
-35
A
B
The location of both distributions is about the same.
Start
Questions
Feedback
System
Manual

References

Frick H, Strobl C, Leisch F, Zeileis A (2012). “Flexible Rasch Mixture Models with Package psychomix.” *Journal of Statistical Software*, **48**(7), 1–25. doi:10.18637/jss.v048.i07

Frick H, Strobl C, Zeileis A (2015). “Rasch Mixture Models for DIF Detection: A Comparison of Old and New Score Specifications.” *Educational and Psychological Measurement*. **75**(2), 208–234. doi:10.1177/0013164414536183

Hothorn T, Zeileis A (2015). “partykit: A Modular Toolkit for Recursive Partytioning in R.” *Journal of Machine Learning Research*, **16**, 3905–3909. http://www.jmlr.org/papers/v16/hothorn15a.html

Kopf J, Zeileis A, Strobl C (2015). “A Framework for Anchor Methods and an Iterative Forward Approach for DIF Detection.” *Applied Psychological Measurement*, **39**(2), 83–103. doi:10.1177/0146621614544195

Kopf J, Zeileis A, Strobl C (2015). “Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches.” *Educational and Psychological Measurement*, **75**(1), 22–56. doi:10.1177/0013164414529792

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

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

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

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

Zeileis A, Umlauf N, Leisch F (2014). “Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond.” *Journal of Statistical Software*, **58**(1), 1–36. doi:10.18637/jss.v058.i01