Detecting Invariance in Psychometric Models with the `psychotree` Package

Carolin Strobl, Florian Wickelmaier,
Julia Kopf and Achim Zeileis

University of Zurich[UZH]

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LMU

universität
innsbruck

Detecting Invariance in Psychometric Models with the psychotree Package

Carolin Strobl, Florian Wickelmaier,
Julia Kopf and Achim Zeileis

presenting: Basil Abou El-Komboz

University of Zurich[UZH]

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LMU

universität innsbruck

Scope of this talk

groups of subjects may

- show higher preferences for certain stimuli
- have an easier time answering certain test items

Scope of this talk

groups of subjects may

- show higher preferences for certain stimuli
- have an easier time answering certain test items
- need separate psychometric model for each group

Scope of this talk

groups of subjects may

- show higher preferences for certain stimuli
- have an easier time answering certain test items
- need separate psychometric model for each group

the aim of this talk is to illustrate

- how parameter instabilities in two widely used psychometric models can be detected
- by means of the `psychotree` package

Example 1: Bradley-Terry trees

- Bradley-Terry models are used for scaling the preferences of subjects for a set of stimuli
- the stimuli and are presented in paired comparisons
- from the subjects' responses the preference scale is estimated

question:

- do all subjects have the same preference scale?

Stimuli: "Germany's next topmodels 2007"

Sample and methods

- sample: $n = 192$ (96 female and 96 male) raters between the age of 15 and 77

- covariates: gender, age and

  (q1) Do you know the women on the photos? Do you know the TV show Germany's Next Topmodel? *

  (q2) Did you watch the latest season of Germany's Next Topmodel regularly?

  (q3) Have you seen the final of the latest season of Germany's Next Topmodel? Do you know who won the latest season of Germany's Next Topmodel? *

  * where "yes" to one or more parts = overall "yes"

- design: forced choice full paired comparison of photos

Algorithm for partitioning psychometric models

- fit a joint model in the starting sample
- select the covariate inducing the strongest parameter instability
- within that covariate: select the cutpoint inducing the highest improvement of model fit
- split the sample
- repeat recursively or stop if
  - no more significant instability
  - sample size too small

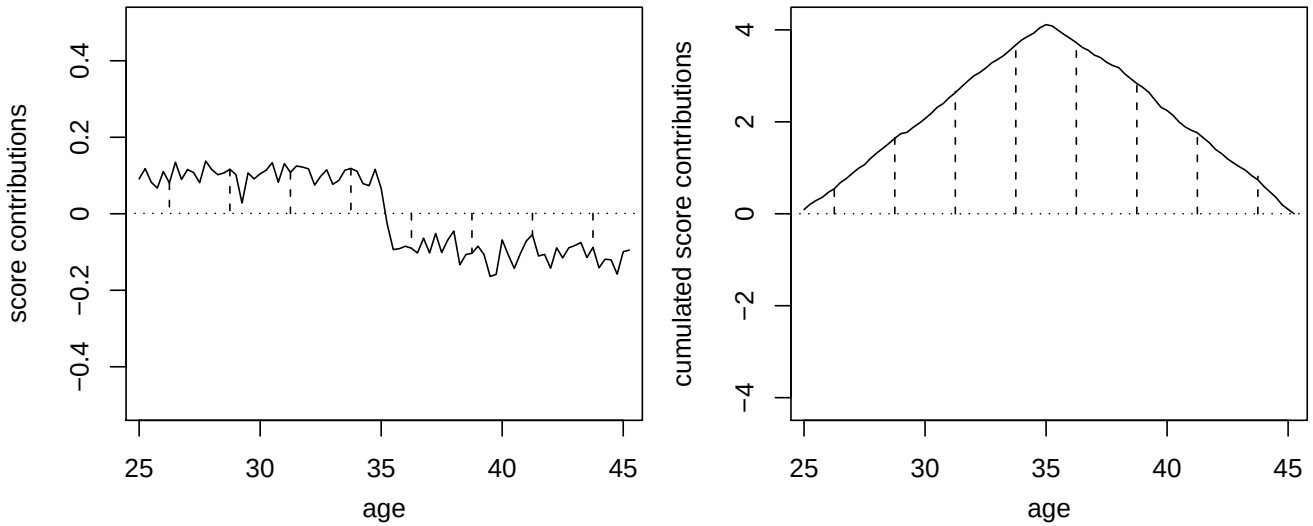
Detect parameter instability

- individual contributions to the score-funktion

$$\psi(y_i, \theta) = \frac{\partial \Psi(y_i, \theta)}{\partial \theta}$$

- cumulated over all values of the covariate ℓ

$$W_\ell(t) = \hat{V}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \psi(y_{(i|\ell)}, \hat{\theta})$$



- under H_0 the path fluctuates randomly around zero ($\rightarrow$ Brownian bridge)

Variable selection

starting with a joint model in the root node

which covariate has the lowest p-value

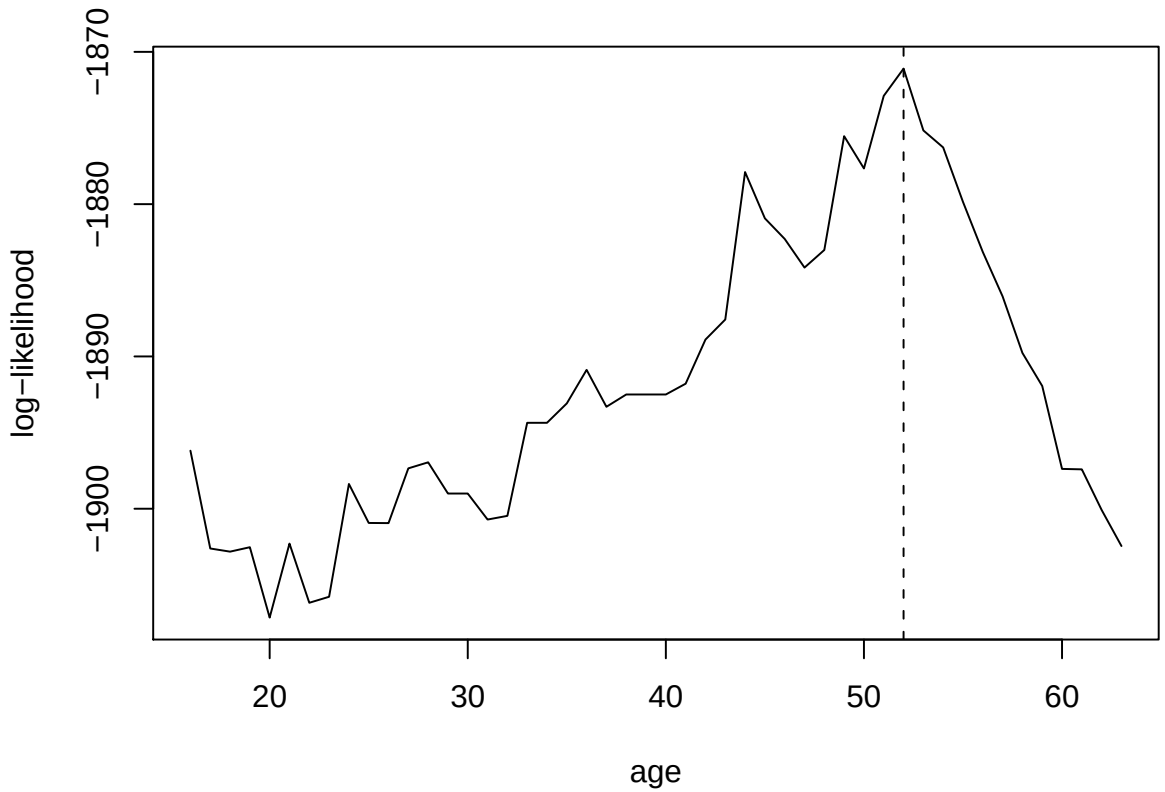
(= induces strongest parameter instability)

	gender	age	q1	q2	q3
statistic	17.0880	32.3566	12.6320	19.8392	6.7586
p-value	0.0217	0.0008	0.1283	0.0067	0.7452

$\Rightarrow$ split in the variable age!

Cutpoint selection

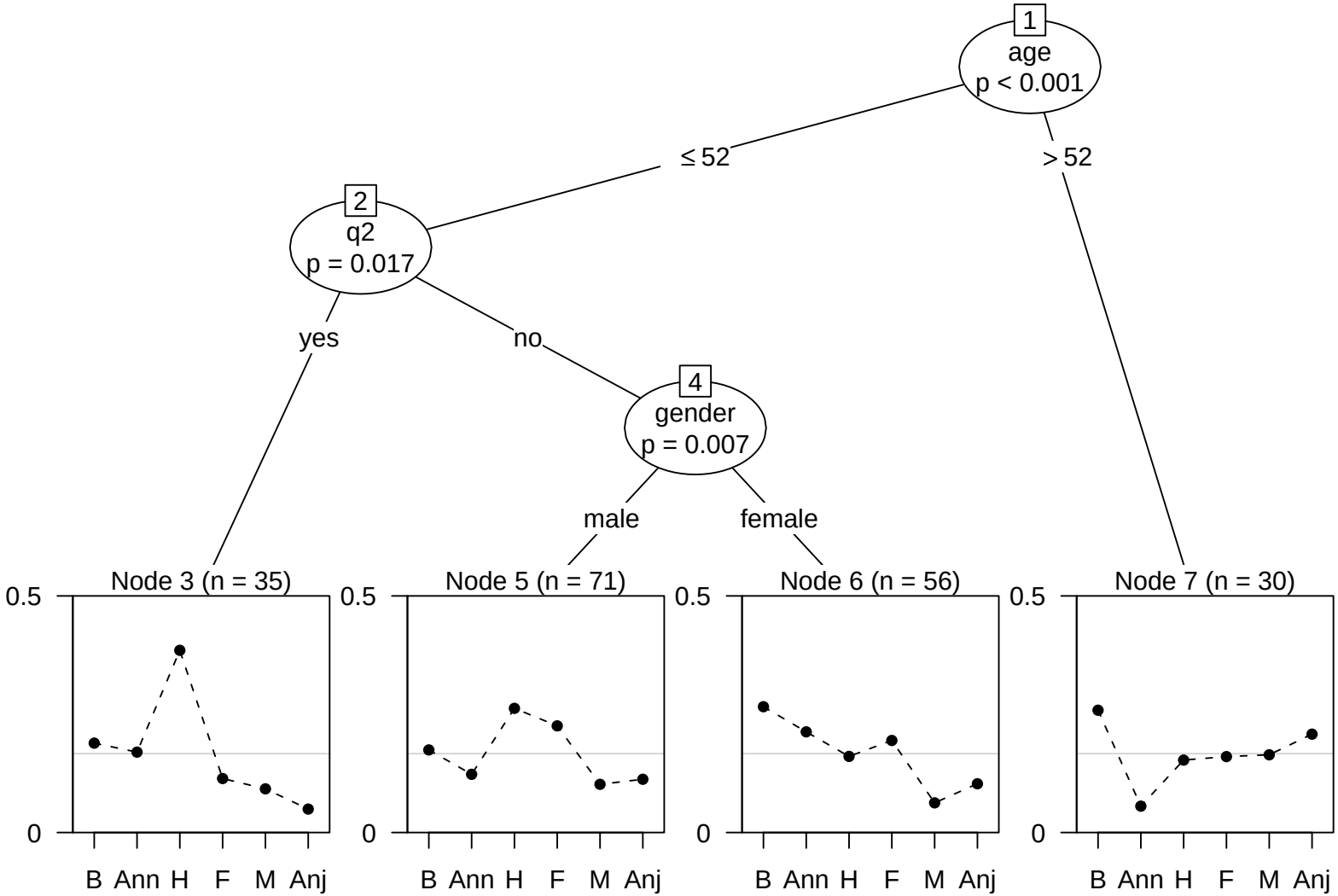
which cutpoint maximizes the partitioned likelihood



$\Rightarrow$ split at the value 52!

keep splitting recursively until stop criterion is reached...

Resulting tree



Example 2: Rasch trees

- the Rasch model is used for measuring latent traits, such as abilities or attitudes
- it contains item parameters and person parameters

to construct a psychological test:

1. estimate item difficulties (via conditional ML)
   $\rightarrow$ detect violations of the model assumptions
2. only if joint model fits: estimate person abilities

question: (refers to step 1)

- are the item difficulties the same for all subjects?

Differential Item Functioning (DIF)

is present when one or more items of a test

- are easier or harder to solve for certain subjects
- even though they have the same ability

Data: "Students-PISA"

- open-access online survey on general education
- conducted by a german weekly news journal
- each partcipant was randomly assigned one of 24 questionnaires, consisting of 45 items from 5 topics: politics, history, economics, culture and natural sciences
- questions were either multiple-choice or open
- recorded response: correct/wrong

results presented here are for one exemplary questionnaire, $n = 9442$

Preliminary (premature) results

if we assume one joint Rasch model and compare the person abilities, we find that those participants who received their high school degree in a certain federal country (Rheinland-Pfalz) perform significantly better

possible explanations:

Preliminary (premature) results

if we assume one joint Rasch model and compare the person abilities, we find that those participants who received their high school degree in a certain federal country (Rheinland-Pfalz) perform significantly better

possible explanations:

- they really have a better general knowledge

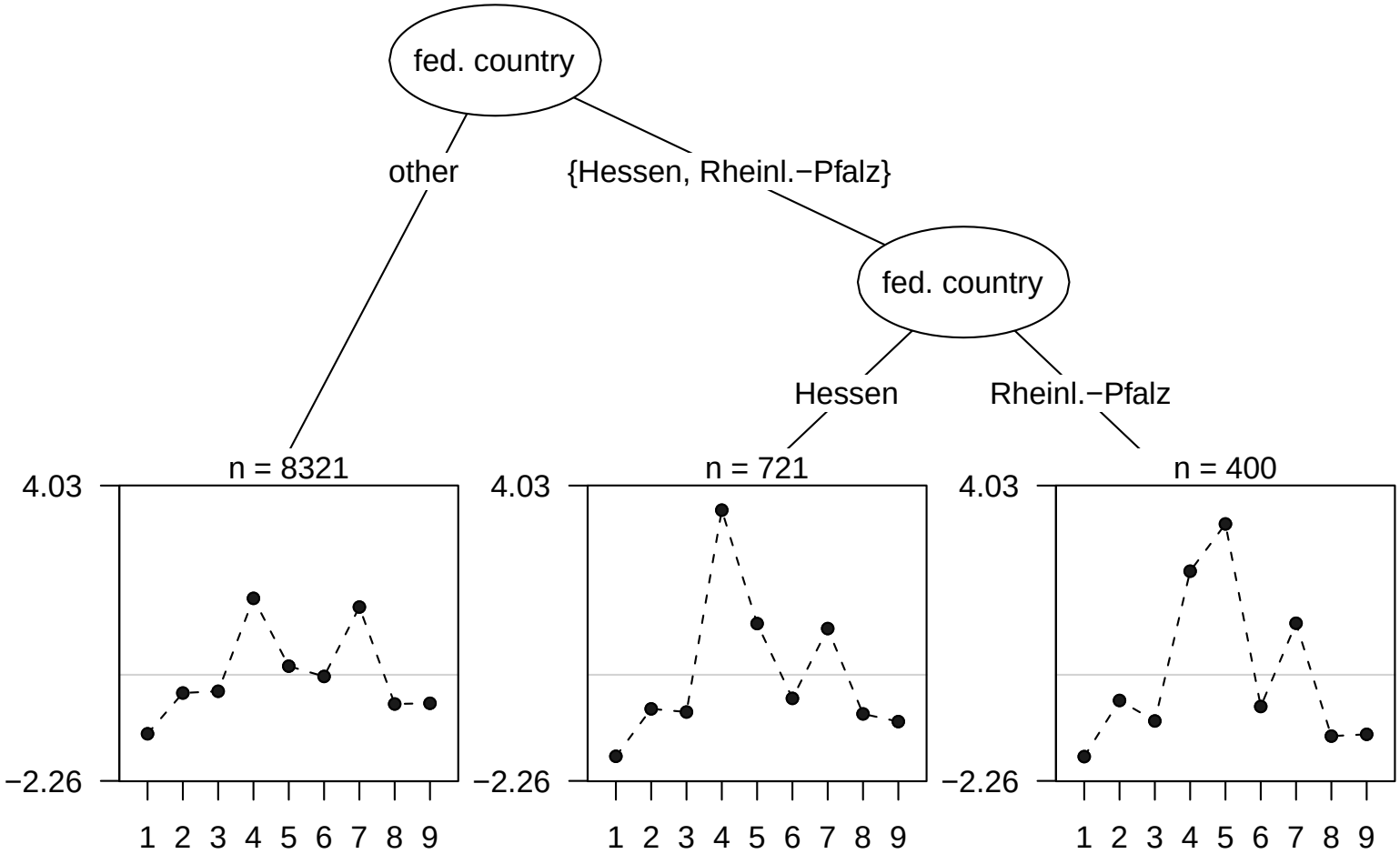
Preliminary (premature) results

if we assume one joint Rasch model and compare the person abilities, we find that those participants who received their high school degree in a certain federal country (Rheinland-Pfalz) perform significantly better

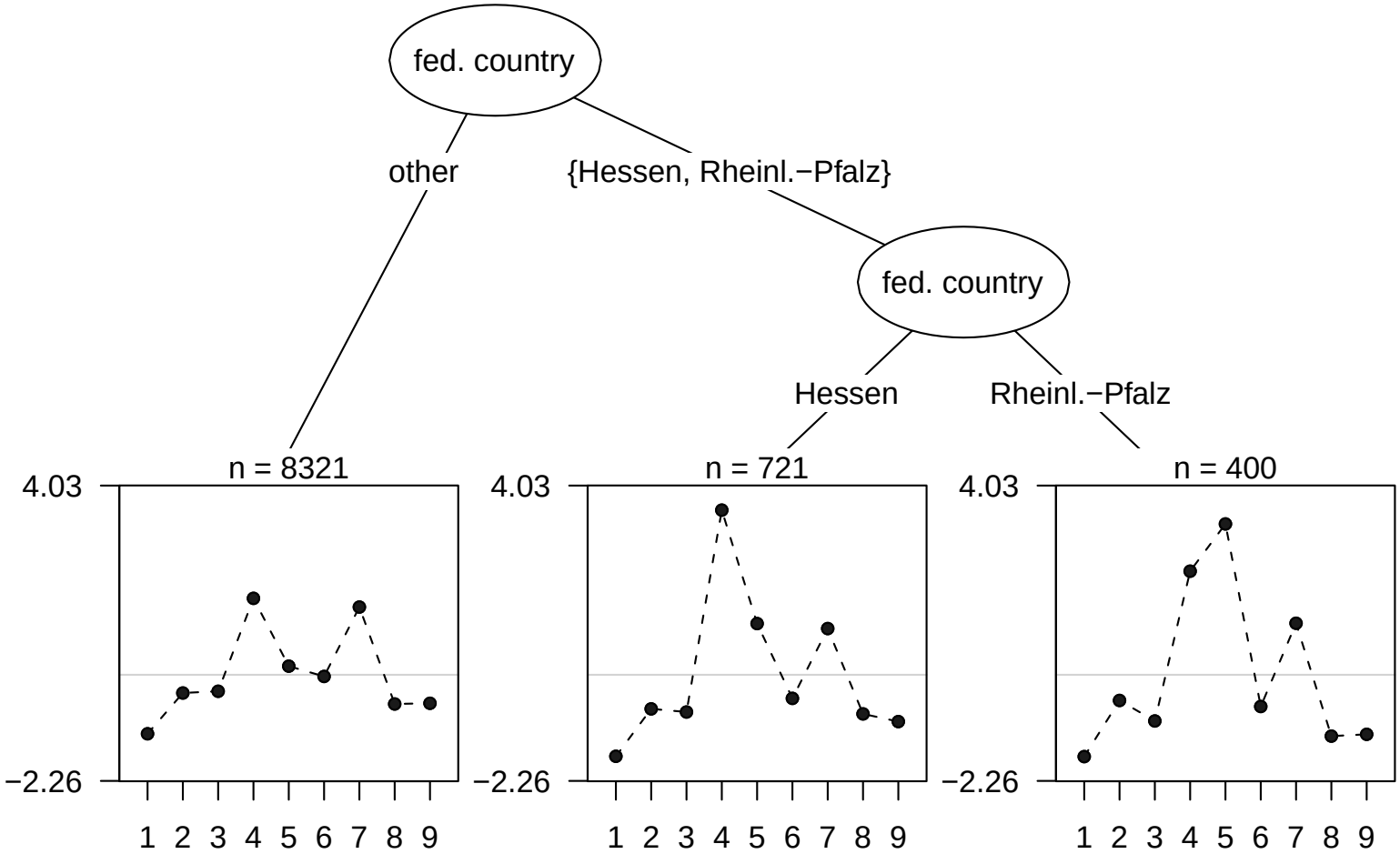
possible explanations:

- they really have a better general knowledge
- they have an unfair advantage $\Rightarrow$ test for DIF

Resulting tree



Resulting tree



Nr. 4: Where is Hessen? (indicate location on a map)

Nr. 5: What is the capital of Rheinland-Pfalz? (Mainz)

Psychological relevance

when DIF is detected

- the test is no longer objective
- fair comparisons between the groups are impossible

$\Rightarrow$ eliminate DIF-items from the test

in our example:

eliminating items 4 and 5 eliminates group differences, i.e.,

Psychological relevance

when DIF is detected

- the test is no longer objective
- fair comparisons between the groups are impossible

$\Rightarrow$ eliminate DIF-items from the test

in our example:

eliminating items 4 and 5 eliminates group differences, i.e.,

the supposed group difference was only an artefact of the test construction!

Package psychotree

- function raschtree for Rasch models

```
raschtree(resp ~ age + gender + motivation,
data = DIFSim)
```

- function bttree for Bradley-Terry models

```
bttree(preference ~ ., data = Topmodel2007,
ref = "Barbara")
```

with respective print, plot and coef methods

- paircomp class for representing and visualizing data from paired comparison experiments
- ongoing work: functions for partitioning polytomous Rasch, two- and three-parameter logistic models

References

Strobl, C., F. Wickelmaier, and A. Zeileis (2010). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics 36*(2), 135–153.

Strobl, C., J. Kopf, and A. Zeileis (2010). A new method for detecting differential item functioning in the Rasch model. *Technical report.* URL `http://epub.ub.uni-muenchen.de`.

Zeileis, A., T. Hothorn, and K. Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics 17*(2), 492–514.

software and documentation available from:

`http://CRAN.R-project.org/package=psychotree`