Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model

Carolin Strobl
Universität Zürich

Julia Kopf
Ludwig-Maximilians-Universität München

Achim Zeileis
Universität Innsbruck

Abstract

A variety of statistical methods have been suggested for detecting differential item functioning (DIF) in the Rasch model. Most of these methods are designed for the comparison of pre-specified focal and reference groups, such as males and females. Latent class approaches, on the other hand, allow to detect previously unknown groups exhibiting DIF. However, this approach provides no straightforward interpretation of the groups with respect to person characteristics. Here, we propose a new method for DIF detection based on model-based recursive partitioning that can be considered as a compromise between those two extremes. With this approach it is possible to detect groups of subjects exhibiting DIF, which are not pre-specified, but result from combinations of observed covariates. These groups are directly interpretable and can thus help generate hypotheses about the psychological sources of DIF. The statistical background and construction of the new method are introduced by means of an instructive example and extensive simulation studies are presented to support and illustrate the statistical properties of the method, that is then applied to empirical data from a general knowledge quiz. A software implementation of the method is freely available in the R system for statistical computing.

Keywords: item response theory, IRT, Rasch model, differential item functioning, DIF, measurement invariance, structural change, model-based recursive partitioning.

1. Introduction

In educational and psychological testing, the term differential item functioning (DIF) 'means that the probability of a correct response among equally able test takers is different for various racial, ethnic, gender [or other] subgroups. A given educational or psychological test consisting of many items with significant DIF may be unfair for certain subgroups, and it is important to identify these items, so that they can be improved or deleted from the test' (Westers and Kelderman 1992).

A variety of statistical methods is available for detecting DIF in the Rasch model. While some of these methods are explicitly designed to detect DIF in individual items, such as the item-specific Wald test (Fischer and Molenaar 1995), others are global goodness-of-fit tests for the Rasch model that also respond to DIF, such as the widely used likelihood ratio test (Andersen 1972; Gustafsson 1980). Most of these methods are based on the comparison of the item parameter estimates between two or more pre-specified groups of subjects, such as males and females as focal and reference groups. This class of model tests also includes the

This is a preprint of an article published in *Psychometrika*, **80**(2), 289–316.
doi:10.1007/s11336-013-9388-3

simple graphical model test as well as the most recent approaches for DIF detection based on a mixed model representation of the Rasch model (Rijmen, Tuerlinckx, De Boeck, and Kuppens 2003; Van den Noortgate and De Boeck 2005).

The advantage of model tests for given groups is that, if DIF is detected, the results can be interpreted straightforwardly in terms of, e.g., which items are easier or harder to solve for which subjects. This can give valuable hints for generating hypotheses about the psychological sources of DIF and how it can be eliminated or avoided in future versions of the test.

On the other hand, in all above-mentioned approaches only those groups that are explicitly proposed by the researcher are tested for DIF. Variables typically proposed for testing include age, gender, ethnicity and language, depending on the objective of the assessment (cf., e.g., Gelin, Carleton, Smith, and Zumbo 2004; Perkins, Stump, Monahan, and McHorney 2006; Woods, Oltmanns, and Turkheimer 2009; Pedraza, Graff-Radford, Smith, Ivnik, Willis, Petersen, and Lucas 2009). However, if in later analyses a group difference is found in a variable that has not been explicitly tested for DIF, it cannot be ruled out that this effect is only an artifact due to unnoticed DIF. Moreover, in most standard approaches numeric variables, such as age, need to be discretized prior to testing, which leads to a loss of information.

At the other extreme, the latent class (or mixture) approach of Rost (1990) tests for item parameter differences between all possible groups of subjects regardless – and even in the absence – of person covariates (see also Kelderman and MacReady 1990; Mislevy and Verhelst 1990). In this sense, the latent class approach can be considered as a very stringent model test (even though it has a lower statistical power than tests for given groups when informative covariates are available, cf. Smit, Kelderman, and Van der Flier 2000). However, the latent class approach provides no straightforward interpretation of the resulting groups. Therefore, often latent class approaches are used only as a first step in the analysis, where the second step is to attempt to describe the latent classes by person covariates for interpretability (see, e.g., Cohen and Bolt 2005; Hancock and Samuelsen 2007; Maij-de Meij, Kelderman, and Van der Flier 2008, and the references therein).

Here, we propose a new statistical approach for detecting DIF in the Rasch model that can be considered as a compromise between the two former approaches – testing only pre-defined and hence easy to interpret groups vs. testing all possible groups in a latent class approach and having to give up interpretability. The idea for the new method is to recursively test all groups that can be defined based on (combinations of) the available covariates – thus preserving interpretability, but still exploring a very wide set of potential indicators of DIF.

In the next section, the rationale and technical details of the new method are first explained by means of a simple artificial example. In Section 3 the results of a series of simulation studies are presented to support and illustrate the statistical properties of the newly proposed method. Finally, an application to empirical data from a general knowledge quiz is presented in Section 4. The proposed method is freely available as a software implementation in the add-on package `psychotree` (Zeileis, Strobl, Wickelmaier, and Kopf 2012) for the R system for statistical computing (R Development Core Team 2012).

2. A new method based on recursive partitioning

The new method for detecting groups of subjects with DIF is based on the technique of model-based recursive partitioning, that employs statistical tests for structural change adopted from

Variable	Summary statistics					
Gender	male: 99			female: 101		
	$x_{\min}$	$x_{0.25}$	x_{med}	$\bar{x}$	$x_{0.75}$	$x_{\max}$
Age	16	30	45	44.27	57	73
Motivation	1	3	4	3.65	5	6

Table 1: Summary statistics for the covariates of the instructive example (artificial data).

econometrics. Model-based recursive partitioning is a semi-parametric approach. The aim is to detect differences in the parameters of a statistical model between groups of subjects defined by (combinations of) covariates.

Model-based recursive partitioning is related to – but by means of modern statistical techniques avoids the earlier weaknesses of – the method of classification and regression trees (CART, Breiman, Friedman, Olshen, and Stone 1984; see Strobl, Malley, and Tutz 2009 for a thorough introduction), where the covariate space is recursively partitioned to identify groups of subjects with different values of a categorical or numeric response variable. As an advancement of this approach, in model-based recursive partitioning it is the parameters of a parametric model – rather than the values of a single response variable – that vary between groups. Such parameters could be, e.g., intercept and slope parameters in a linear regression model or, as in our case, the item parameters of a Rasch model that may vary between groups of subjects.

This principle is now first illustrated by means of an artificial instructive example, before the technical details are addressed in the next sections. The data for the instructive example consist of the simulated responses of 200 subjects to 20 items, which can be considered, e.g., as questions in a proficiency test. In addition to the responses, the data set includes three covariates: gender, age, and a motivation score. The summary statistics for the latter are listed in Table 1.

The data for the instructive example were simulated with DIF between three groups: males up to the age of 35, males above the age of 35, and females. Item 3 was simulated to be more difficult for women and younger men, item 11 was simulated to be more difficult only for women and items 14 and 15 were simulated to be easier only for younger men. (These items are highlighted in Figure 1 for illustration.) No DIF was generated in the variable motivation.

In order to detect DIF with the new method, the item responses are assessed with respect to possible group differences related to the three covariates gender, age, and motivation, as described in detail below. The resulting model, that is partitioned with respect to a combination of the covariates gender and age, is presented in Figure 1 and will be termed a Rasch tree from here on. In each of the terminal nodes of the tree, the item parameter estimates for the 20 items are displayed (a high value indicates a high difficulty of the item).

Following the tree from top to bottom, we find that different item parameters result for males and females, and within the group of males for those up to the age of 35 and over the age of 35. For example, the third item is harder for males up to the age of 35 (represented in node 3) and females (represented in node 5) than for males over the age of 35 (represented in node 4).

Generally speaking, the mere fact that there is more than one terminal node in Figure 1 means that the null hypothesis of one joint Rasch model for the entire sample (i.e. measurement

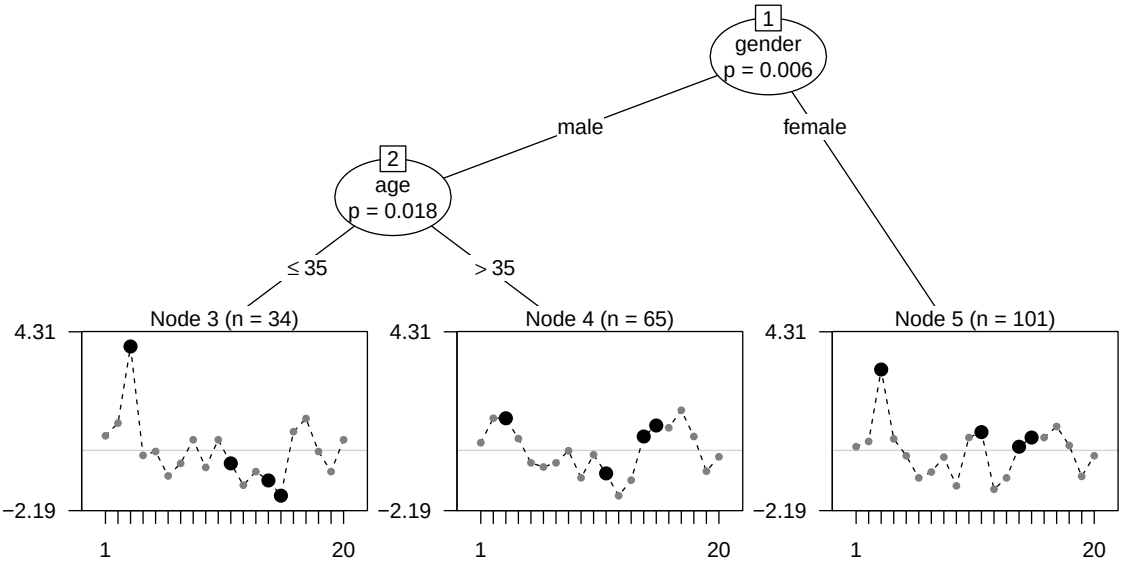

Figure 1: Rasch tree for the instructive example (artificial data for illustration purposes), exhibiting DIF between males up to the age of 35, males over the age of 35, and females. In the terminal nodes, estimates of the item difficulty are displayed for each of the 20 items.

invariance) must be rejected. In this sense, the proposed method is a global test for DIF as well as an overall model test for the Rasch model. In addition to this, we can see from the graphical visualization which groups are affected by DIF with respect to which items. This information can help generate hypotheses about the underlying sources of DIF and guide the decision how to proceed with the affected items.

Figure 1 also shows that the simulated pattern of covariates associated with DIF was correctly replicated by the Rasch tree. In particular, the fact that some item parameters differ between males up to the age of 35 and males above the age of 35 was correctly discovered by the Rasch tree. As opposed to that, the widely employed approach of arbitrarily splitting a numeric variable at the median (which in this case would have been at the value 45 and thus far too high) would not only conceal the actual age at which the parameter change occurs but may even result in not detecting significant DIF in a numeric variable at all, as is further illustrated in the simulation studies below. Moreover, the variable motivation was not selected for splitting (i.e. no DIF was detected with respect to motivation), which also correctly replicates the simulated pattern.

What is important to note here is that the entire structure identified by the Rasch tree – i.e. that the DIF groups are formed by this particular combination of the two variables gender and age, including the location of the cutpoint in the variable age – was not pre-specified and provided to the algorithm, but was learned from the data in an exploratory way. This is a key feature of the model-based recursive partitioning approach employed here, that makes it very flexible for detecting groups with DIF and distinguishes it from parametric regression models, where only those main effects and interactions that are explicitly included in the specification of the model are considered.

Technically, the following consecutive steps are used to infer the structure of a Rasch tree like that depicted in Figure 1 from the data:

1. Estimate the item parameters jointly for all subjects in the current sample, starting with the full sample.
2. Assess the stability of the item parameters with respect to each available covariate.
3. If there is significant instability, split the sample along the covariate with the strongest instability and in the cutpoint leading to the highest improvement of the model fit.
4. Repeat steps 1–3 recursively in the resulting subsamples until there are no more significant instabilities (or the subsample becomes too small).

These four steps are now explained in more detail.

2.1. Estimating the item parameters

We use the common conditional maximum likelihood approach for estimating the item parameters (but the method can in principle also be adapted to other maximum likelihood estimation approaches). Let θ_i, $i = 1, \ldots, n$, denote the person parameters, β_j, $j = 1, \ldots, m$, denote the item parameters and u_{ij} denote the response of subject i to item j. Since under the Rasch model

$$P(U_{ij} = u_{ij}|\theta_i, \beta_j) = \frac{e^{u_{ij} \cdot (\theta_i - \beta_j)}}{1 + e^{\theta_i - \beta_j}}$$

the person raw-scores $r_i = \sum_{j=1}^{m} u_{ij}$ form sufficient statistics for the person parameters, the item parameters can be estimated by means of iterative procedures from the conditional likelihood

$$L_c(\boldsymbol{\beta}|r_1, \ldots, r_n) = \prod_{i=1}^{n} L_c(\boldsymbol{\beta}|r_i) = \prod_{i=1}^{n} \frac{e^{-\sum_{j=1}^{m} u_{ij} \cdot \beta_j}}{\gamma_{r_i}(\boldsymbol{\beta})}, \tag{1}$$

where γ_{r_i} is the symmetric function of order r_i (cf., e.g., Fischer and Molenaar 1995). To fix the origin of the scale, some constraint has to applied, e.g., setting the first item parameter or the sum of all item parameters to zero, leaving $m - 1$ free parameters.

2.2. Testing for parameter instability

In order to test whether the item parameters vary between groups of subjects defined by covariates, we use the approach of structural change tests from econometrics. These tests are usually employed for detecting, e.g., a drop in stock returns over time, whereas here we employ the same methodology for detecting parameter changes over person covariates.

The rationale of the employed structural change tests is the following: The item parameters are first estimated jointly for the entire sample. Then the individual deviations from this joint model are ordered with respect to a covariate, such as age. If there is systematic DIF with respect to groups formed by the covariate, the ordering will exhibit a systematic change in the individual deviations. If, on the other hand, no DIF is present, the values will merely fluctuate randomly.

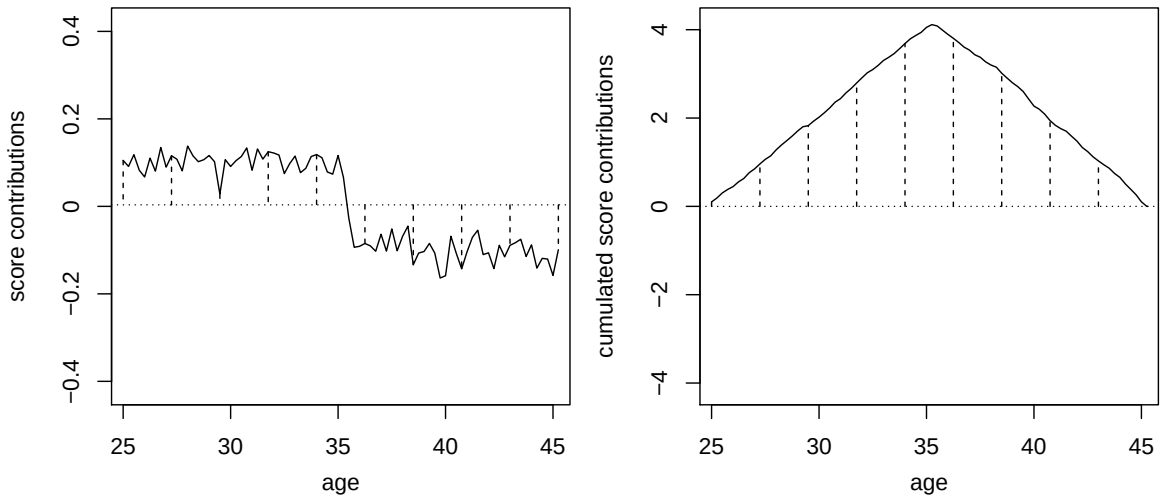

Figure 2: Structural change in the variable age (artificial data for illustration purposes). In the left plot, the individual score contributions are ordered with respect to the variable. The dashed lines indicate deviations from the overall mean zero, which are positive before the structural change and negative afterwards. In the right plot, the positive and negative deviations are cumulated and the structural change is now noticeable from the peak in the cumulative sum process.

This rationale is illustrated in Figure 2: The individual contributions of all subjects to the score function, that is used for the estimation of a parameter (details follow below), are ordered with respect to the variable age, as visualized in the left hand side of Figure 2. By definition, the score contributions are zero on average. However, when the score contributions are ordered with respect to the variable age, it becomes obvious that they do not fluctuate randomly around the mean zero – which would be the case under the null hypothesis that one joint parameter estimate is appropriate for the entire sample – but there is a systematic change at age 35. This systematic change indicates that, instead of one joint parameter estimate for the entire sample, different parameter estimates should be permitted for subjects up to the age of 35 and above the age of 35.

For statistically testing structural change in the model parameters, we suggest the usage of generalized M-fluctuation tests (Zeileis and Hornik 2007) that form the basis of the model-based recursive partitioning framework of Zeileis, Hothorn, and Hornik (2008). The idea of this class of tests is to compute the subject-wise score contributions (i.e. the deviations from a joint model, that are illustrated in the left hand side of Figure 2) and derive test statistics with known distributions from them.

The individual score function $\psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}})$, for $i = 1, \ldots, n$ observations, i.e., the derivative of the individual contributions to the log-likelihood $\Psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}})$ with respect to the parameter vector, is a general measure of deviation for likelihood-based models. For the Rasch model these individual contributions can easily be computed from the conditional likelihood as outlined below.

For the construction of the test statistic, the individual contributions to the score function

are cumulated according to the order induced by the variable age, as illustrated in Figure 2, or any other covariate. The systematic change from positive to negative in the individual contributions to the score function in the left hand side of Figure 2 is then captured as a clearly noticeable peak in the cumulative sum process in the right hand side of Figure 2.

The cumulative sum process is defined as

$$W_\ell(t) \quad = \quad \widehat{\boldsymbol{V}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \psi(\boldsymbol{u}_{(i|\ell)}, \hat{\boldsymbol{\beta}}) \qquad (0 \leq t \leq 1), \tag{2}$$

where the index $(i|\ell)$ denotes the i-th ordered observation with respect to the ℓ-th covariate, $\lfloor \cdot \rfloor$ denotes the integer part, $\widehat{\boldsymbol{V}} = \sum_{i=1}^{n} \psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}) \psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}})^\top$ is the outer-product-of-gradients estimate of the covariance matrix, and t is a fraction of the sample size. Under the null hypothesis of parameter stability, the cumulative sum process $W_\ell(\cdot)$ can be shown to converge to an $(m-1)$-dimensional Brownian bridge (Zeileis and Hornik 2007), which can be used as the basis for statistical inference.

The cumulative aggregation runs over the order induced by the ℓ-th covariate: The $i = 1, \ldots, n$ individual deviations are ordered with respect to the covariate and aggregated up to the $\lfloor n \cdot t \rfloor$-th element in each step. When $W_\ell(t)$ is considered as a function of the fraction t of the sample size, under the null hypothesis of parameter stability the cumulative sum process follows the path of a random process with constant zero mean (whereas under the alternative hypothesis of parameter instability the path deviates from this random fluctuation, as illustrated in the right hand side of Figure 2).

The advantage of this approach is that the model does not have to be reestimated for all splits in all covariates, because the individual deviations remain the same and only their ordering (and the corresponding path of $W_\ell(t)$) needs to be adjusted for evaluating the different covariates.

To capture systematic deviations in $W_\ell(\cdot)$, different test statistics can be used depending on whether the ℓ-th covariate is a numeric or a categorical variable. If it is numeric, Zeileis *et al.* (2008) point out that a natural test statistic is

$$S_\ell \quad = \quad \max_{i=\underline{i},\ldots,\overline{i}} \left(\frac{i}{n} \cdot \frac{n-i}{n} \right)^{-1} \left\| W_\ell\left(\frac{i}{n}\right) \right\|_2^2. \tag{3}$$

This can be interpreted as the maximum Lagrange multiplier statistic (also known as score statistic) for a single shift alternative over all conceivable cutpoints in $[\underline{i}, \overline{i}]$. The limiting distribution is the supremum of a tied-down Bessel process, from which p values can be computed (for details see Zeileis *et al.* 2008; Merkle and Zeileis 2013).

If, on the other hand, the ℓ-th covariate is categorical (with values $x_{i\ell}$ taking categories $q = 1, \ldots, Q$), it is more natural to use the following test statistic

$$S_\ell \quad = \quad \sum_{q=1}^{Q} n \left(\sum_{i=1}^{n} I(x_{i\ell} = q) \right)^{-1} \left\| \Delta_q W_\ell\left(\frac{i}{n}\right) \right\|_2^2, \tag{4}$$

where Δ_q is the increment within the q-th category. This test statistic is invariant to reordering of the Q categories and the subjects within each category. The test statistic captures the instability over the Q subsamples. Its limiting distribution is χ^2 with $(Q-1) \cdot (m-1)$

		Node 1	Node 2	Node 3	Node 4	Node 5
Age	Statistic	41.237	48.448	28.924	37.699	25.678
	p value	0.171	0.018*	0.593	0.208	0.961
Gender	Statistic	41.479	—	—	—	—
	p value	0.006*	—	—	—	—
Motivation	Statistic	112.368	94.680	84.078	105.762	120.598
	p value	0.290	0.740	0.432	0.378	0.077

Table 2: Summary of the parameter instability test statistics and corresponding Bonferroni adjusted p values for the instructive example. Those variables whose p values are highlighted with * symbols are selected for splitting in the respective node.

degrees of freedom, from which p values can be computed. This test is employed for both nominal and ordinal categorical variables. A potential ordering of the categories is accounted for in the next step, when the cutpoint is selected (see Section 2.4 below).

For the Rasch model, the objective function used for parameter estimation is the conditional log-likelihood. The individual contributions to the conditional log-likelihood can be easily computed as $\log L_c(\boldsymbol{\beta}|r_i)$ (cf. Equation 1), yielding

$$\Psi(\boldsymbol{u}_i, \boldsymbol{\beta}) = -\sum_{j=1}^{m} u_{ij} \cdot \beta_j - \log\left(\gamma_{r_i}(\boldsymbol{\beta})\right). \tag{5}$$

For the computation of the structural change tests, the individual contributions to the score function are derived from Equation 5. The contribution of the i-th subject for the j-th item parameter is:

$$\psi(\boldsymbol{u}_i, \boldsymbol{\beta})_j = \frac{\partial \Psi(\boldsymbol{u}_i, \boldsymbol{\beta})}{\partial \beta_j} = -u_{ij} - \frac{1}{\gamma_{r_i}(\boldsymbol{\beta})} \cdot \frac{\partial \gamma_{r_i}(\boldsymbol{\beta})}{\partial \beta_j} \tag{6}$$

The derivatives of the symmetric functions $\gamma_{r_i}(\boldsymbol{\beta})$ are again symmetric functions with certain terms omitted (cf., e.g., Fischer and Molenaar 1995). In our implementation of the Rasch tree, the sum algorithm of Liou (1994) is used (by default) for computing these derivatives.

When the individual contributions to the score function of the Rasch model from Equation 6 are ordered with respect to covariate ℓ and inserted in Equation 2, parameter instabilities in the item parameters can be statistically tested using the model-based recursive partitioning approach outlined above.

The results of this procedure are also easy to interpret: The parameter instability test statistics S_ℓ with associated (Bonferroni adjusted, cf. Section 2.5) p values are provided for each candidate variable, as illustrated for the instructive example in Table 2. The test statistics correspond to Equation 3 for the numeric variable age and to Equation 4 for the categorical variable gender and the ordered categorical variable motivation. The p values are derived from the respective limiting distributions.

In the first node, the variable with the smallest p value – in this case gender – is selected for splitting (cf. Table 2 and Figure 1). In each daughter node the splitting continues recursively: Here, the variable age is selected for splitting in the second node, whereas no further splits are found significant in the following nodes.

Note that the variable gender is no longer available for splitting after the first node because it offers only one possible cutpoint (that has already been used for the first split). As opposed

to gender, the second splitting variable age offers as many possible cutpoints as it has distinct values. In this case, it is an important advantage of the model-based recursive partitioning method that the exact cutpoint does not need to be pre-specified, but is determined in a data-driven way as described in detail in Section 2.4.

Splitting continues until all p values exceed the significance level (commonly 5%), indicating that there is no more significant parameter instability, or until the number of observations in a subsample falls below a given threshold.

2.3. Computational aspects

The model-based recursive partitioning approach outlined here employs a Lagrange multiplier (LM) or score test – rather than, e.g., a likelihood ratio (LR) or Wald test, that are equally well established for the Rasch model (cf., e.g., Fischer and Molenaar 1995, Chapter 5) – in the variable selection step. One reason for this is the general construction of the test statistic and the resulting differences in the computational burden.

When the LR test is used to test, e.g., whether two or more groups have different item parameters, the parameters need to be estimated both for the full sample and for all subsamples. The full sample likelihood for the full sample item parameter estimates is then compared to the product of all subsample likelihoods for the subsample item parameter estimates. For the Wald test, on the other hand, the item parameters need to be estimated for all subsamples only. The subsample item parameter estimates are then directly compared, so that the Wald test does not require computation of the full sample item parameter estimates. Finally, the LM or score test employs only the item parameter estimates from the full sample, and evaluates group differences by means of the individual score contributions, as illustrated above. Asymptotically all three tests are equivalent and hence, in practice, the choice of test is often guided by computational considerations: The LR test is often found to perform slightly better in finite samples; however, it also poses the highest computational burden (see also Merkle and Zeileis 2013, who discuss similarities and differences of the three types of tests in more detail in a structural change setting).

Therefore, from a computational point of view, basing the variable selection decision of the model-based recursive partitioning algorithm on a LM statistic has two advantages:

Even for given groups, the LM test is computationally more efficient than either LR or Wald test. More importantly, however, when the groups are not given a priori, but partitions of the data based on different covariates are investigated, like in the model-based recursive partitioning approach presented here, the LM test has the great advantage that the item parameters have to be estimated only once for the current sample, and tests for all covariates can be constructed simply by re-ordering the individual score contributions, as outlined above.

Consequently, despite the exploratory exhaustive search character of the model-based recursive partitioning algorithm, the computational burden for each variable selection decision is much lower than one might expect.

Besides these considerations of computational complexity, it should be noted that for evaluating DIF in the Rasch model both the LM test employed here and the widely known LR test suggested by Andersen (1972) follow the same principles: Both are global – as opposed to item-wise – tests for DIF based on the conditional likelihood. The LR test was first suggested by Andersen (1972) as a test for different slope parameters by means of dividing the subjects into groups according to their ability raw scores. However, it has long been noted (e.g. by

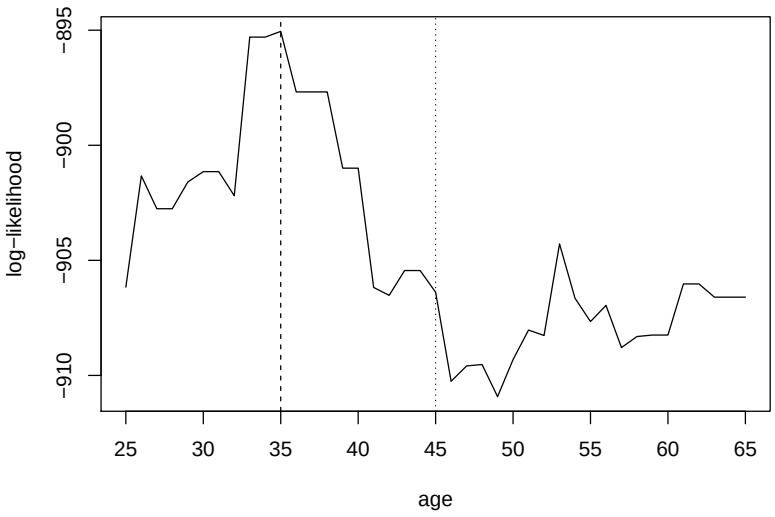

Figure 3: Log-likelihood of the partitioned Rasch model for the second split in the covariate age. The dashed line indicates the location of the optimal cutpoint (at the value 35) while the dotted line indicates the location of the median (at the value 45).

Gustafsson 1980) that it can also be used as a test for DIF (that was still referred to as "item bias" by Gustafsson 1980) when the subjects are divided into groups according to covariates such as gender and social background. Hence, one could easily use the maximum LR statistic (rather than the maximum LM statistic) in Equation 3 as it has the same asymptotic distribution. However, this would give up the computational advantages of the LM test outlined above.

2.4. Selecting the cutpoints

After a covariate has been selected for splitting, the optimal cutpoint is determined by maximizing the partitioned log-likelihood (i.e., the sum of the log-likelihoods for two separate models: one for the observations to the left and up to the cutpoint, and one for the observations to the right of the cutpoint) over all candidate cutpoints within the range of this variable.

For the first split in the instructive example, the selection of the cutpoint is trivial – since the binary variable gender only allows for a single split between the subgroups of females and males. In the second split, however, all possible cutpoints in the variable age for the male subsample are considered and the associated partitioned log-likelihood is displayed in Figure 3. The value 35 is selected as the optimal cutpoint, because it shows the highest value of the partitioned log-likelihood, i.e., the strongest difference in the item parameters exists between males up to the age of 35 and over the age of 35.

Note that other potential cutpoints close to this value also show a high value of the partitioned log-likelihood, so that in different random samples from the same data generating process not always the exact same value for the optimal cutpoint may be detected. However, from

Figure 3 it is obvious that the median (dotted line), that is often used for pre-specifying the focal and reference group from a numeric predictor variable, may be far off the maximum of the partitioned log-likelihood indicating the strongest parameter change. As opposed to that, the data-driven approach suggested here can not only reliably detect the parameter instability in the variable age, but it can also identify at what age the strongest parameter change occurs.

Formally, for a numeric splitting variable ℓ with values $x_{i\ell}$ we can define the subsamples $L(\xi) = \{i \,|\, x_{i\ell} \leq \xi\}$ and $R(\xi) = \{i \,|\, x_{i\ell} > \xi\}$ on the left and right, respectively, of some cutpoint ξ. For both subsamples, the parameters $\hat{\boldsymbol{\beta}}^{(L)}$ and $\hat{\boldsymbol{\beta}}^{(R)}$ can be estimated separately as described above. To determine the optimal cutpoint ξ, the partitioned log-likelihood

$$\sum_{i \in L(\xi)} \Psi\left(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}^{(L)}\right) + \sum_{i \in R(\xi)} \Psi\left(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}^{(R)}\right)$$

is maximized over all candidate cutpoints ξ (typically requiring a certain minimal subsample size).

While this approach can be applied to numeric and ordered covariates, for unordered categorical covariates the Q categories can be split into any two groups. From all these candidate binary partitions, again the one that maximizes the partitioned log-likelihood is chosen.

Note that choosing the optimal cutpoint by maximizing the partitioned (log-)likelihood corresponds directly to using the maximum LR statistic of the joint vs. the partitioned model. Thus, for selecting the optimal cutpoint the computationally more expensive LR test is implicitly used in the Rasch tree algorithm. However, it is not employed in the first step for *testing whether* there is significant DIF in a covariate, but only for the second step of *estimating where* the strongest DIF occurs by obtaining the maximum likelihood estimator for the cutpoint. Unlike the tests in the previous sections, this computationally costly LR test is not applied to all potential splitting variables but only to those selected for splitting in the first place.

From a statistical point of view, this two-step approach – where the variable selection is made independently from the cutpoint selection – has two important advantages: Not only does it considerably reduce the computational burden (as is also illustrated in the following simulation studies), but at the same time it also prevents an artefact termed variable selection bias (cf., e.g., Dobra and Gehrke 2001; Shih 2004; Hothorn, Hornik, and Zeileis 2006; Strobl, Boulesteix, and Augustin 2007), that was inherent in earlier recursive partitioning algorithms.

Variable selection bias occurs when first the best cutpoint is determined in each variable and then the best splitting variable is selected by means of evaluating some splitting criterion or test statistic, that was computed exactly for the cutpoint producing the highest value of this criterion or statistic. In this case, variables offering more cutpoints (such as numeric variables or variables with many categories) have a higher chance of being selected only due to multiple testing, which does not reflect the actual quality of the splitting variable. Therefore, with respect to selecting the variable with the strongest parameter instability in the Rasch model it would be statistically incorrect to select the best splitting variable by means of the standard LM or LR test (based on the standard χ^2 distribution) when the test statistic is computed in the best cutpoint offered by that variable. The reason is that – due to the optimal selection of the cutpoint – the asymptotic distribution of this optimally selected statistic is no longer χ^2 (Andrews 1993). Therefore, the correct distribution has to be derived for any

optimally selected statistic (cf., e.g., Miller and Siegmund 1982; Koziol 1991; Hothorn and Lausen 2003; Boulesteix 2006) – as in our case for the optimally selected LM statistic from Equation 3, that is employed in the variable selection step of the Rasch tree method. This approach guarantees that the selection of the best splitting variable is not affected by the number of cutpoints offered by each candidate variable, but can make the test decision a little conservative in small to moderate samples, as we will see in some of the following simulation studies.

2.5. Stopping criteria

For creating a Rasch tree, the four basic steps outlined above – (1) estimating the item parameters of a joint model, (2) testing for parameter instability, (3) selecting the splitting variable and cutpoint and (4) splitting the sample accordingly – are repeated recursively until a stopping criterion is reached.

Two kinds of stopping criteria are currently implemented: Splitting continues only as long as significant parameter instability is detected. If there is no (more) significant instability with respect to any of the covariates, the splitting stops, as was illustrated in Table 2. Thus, the significance level – usually set to 5% – serves as the most important stopping criterion.

In addition to that, as a second stopping criterion a minimum sample size per node can be specified. This minimal node-size should be chosen such as to provide a sufficient basis for parameter estimation in each subsample, and should thus be increased when the number of item parameters to be estimated is large. For all our examples, a significance level of 5% and a minimal node-size of 20 were employed.

Finally, one should keep in mind that when a large number of covariates is available in a data set, and all those covariates are to be tested for DIF, multiple testing becomes an issue – as with any statistical test for DIF. To account for the fact that multiple testing might lead to an increased false-positive rate when the number of available covariates is large, a Bonferroni adjustment for the p value splitting criterion is applied internally (so that all p values reported for Rasch trees throughout this paper have already been Bonferroni-corrected unless explicitly stated otherwise).

Another issue related to stopping criteria in recursive partitioning algorithms is their potential for overfitting: In classical algorithms (such as CART; Breiman *et al.* 1984) a pruning step (i.e. cutting back branches at the bottom of the tree that do not add to the prediction accuracy in cross-validation) is necessary to make sure that any splits detected for the learning data do not only reflect random variation but also generalize to other samples from the same data generating process. As opposed to these classical algorithms, the model-based recursive partitioning approach employed here is already based on statistical inference tests (rather than merely descriptive statistics) and uses their p values (together with several precautions against multiple testing) for stopping before overfitting occurs (cf. also Hothorn *et al.* 2006). Therefore, pruning is not necessary in this approach.

Moreover, it is important to note that our model-based recursive partitioning algorithm is not affected by an inflation of chance due to its recursive nature. Indeed, several statistical tests are successively conducted in a Rasch tree – but each test is conducted only if the previous test yielded a significant result. In this sense, the recursive approach forms a closed testing procedure, which does not lead to an inflation of chance as is well known from the literature on multiple comparisons (Marcus, Peritz, and Gabriel 1976; Hochberg and Tamhane 1987).

For the Rasch tree this means that the postulated significance level holds for the entire tree, not only for each individual split. This ensures that DIF is not erroneously detected as an artefact of the recursive nature of the algorithm.

These statistical properties are now further illustrated in a series of simulation studies.

3. Simulation studies

The following simulation studies were conducted to empirically support and illustrate the statistical properties of the newly suggested Rasch tree method and compare it to the behavior of the established LR test for given groups.

All simulations were conducted in the R system for statistical computing (R Development Core Team 2012), using our own add-on package `psychotree` (Zeileis *et al.* 2012) for the Rasch tree and the add-on package `eRm` (Mair and Hatzinger 2007; Mair, Hatzinger, and Maier 2012) for the LR test. Further information on software and documentation is provided in the section on computational details at the end of the paper.

3.1. Criterion variables

In order to evaluate whether each method correctly captured the data generating process and to assess the computational effort, the following criterion variables were recorded in each simulation study:

- Percentage of significant test results

  Under the null hypothesis scenarios, where no DIF is simulated in the data generating process, the percentage of significant test results reflects the type I error rate of the method.

  Under the alternative scenarios, where DIF is simulated in the data generating process, the percentage of significant test results reflects the statistical power of the method.

- Root mean squared error (RMSE) of parameter estimation

  The RMSE is computed as the root mean squared difference between the true (simulated) and the estimated parameter for the third item, in which DIF is simulated in the alternative scenarios. Its value thus indicates how well the simulated DIF is recovered by each method. In particular, a larger RMSE is expected in situations where the true (simulated) group structure is not recovered.

  Therefore, the RMSE is of interest for comparing the performance of the methods only in those scenarios where DIF is present and the groups are not entirely pre-defined. Thus, for readability, the RMSE is only presented in the alternative scenarios with DIF for simulation studies I and III.

- Adjusted Rand index (ARI) of group recovery

  The adjusted Rand index (ARI, Hubert and Arabie 1985; Milligan and Cooper 1986) is a measure for the agreement between two partitions of a data set. It is commonly used for comparing the results of cluster analyses to each other or to the true class membership, and is adjusted for agreement by chance.

Here, the ARI is used to measure how well the true (simulated) reference and focal groups are recovered by each method: If the agreement between the true and the recovered partition is high, the ARI shows a high value up to the maximum of 1. If the agreement is poor, the ARI shows a lower value. In particular, the ARI shows the value 0 in cases where, e.g., two distinct reference and focal groups are simulated but only a single group is recovered by the method – i.e. in cases where simulated DIF is not detected.

Therefore, the ARI is also of interest only in those scenarios where DIF is present and the groups are not entirely pre-defined. Thus, for readability, the ARI is only presented in the alternative scenarios with DIF for simulation studies I and III as well.

The recovered partition for computing the ARI is derived in the following way: For the LR test the recovered partition corresponds to the specified reference and focal groups when the test shows significant DIF, and to one single group when no significant DIF is detected. For the Rasch tree the recovered partition directly corresponds to the terminal nodes of the trees, and – like for the LR test – one single group results when no significant DIF is detected.

- Bias, variance and mean squared error (MSE) of cutpoint estimation

  In order to assess the quality of the cutpoint estimation, which is an important aspect of the group recovery, an additional analysis displaying the bias, variance and MSE of the cutpoint estimation for a numeric covariate is conducted for simulation study I.

- Computation time

  The average computation time for one replication in seconds is reported as an indicator of the computational complexity of the method. Since the distribution of the computation times can be very skewed for Rasch tree (especially in those scenarios where the optimal cupoint needs to be selected for a numeric variable), not only the mean but also the median and maximum computation times are reported.

  For the LR test the computation time was recorded for the testing step only (not for the estimation of the Rasch model), while for the Rasch tree the entire procedure (including the estimation of the Rasch model) was timed – which is an agreement in favor of the LR test. Note, however, that any differences in the computation times may indicate differences in the implementations rather than theoretical differences between the two methods.

  Computation times for the simulation studies were recorded on a multiprocessor system with 4 AMD Opteron 6174 processors with 2.2GHz and 12 cores each. Thus, computation times on a new laptop or desktop computer with a more powerful processor may be expected to be faster than the ones reported here, as illustrated for the application example in Section 4.

3.2. Experimental settings

A range of experimental factors was varied in each simulation study, as described in detail for each study below. The following settings were the same for all experiments:

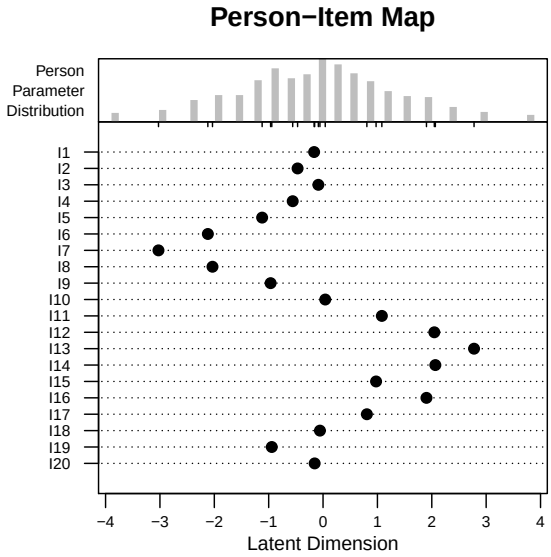

Figure 4: Person-item-map for the simulated data under the null hypothesis scenario.

- Number of replications

  5000 replications were conducted for each experimental scenario to ensure an appropriate precision of the estimates for type I error and power.

- Number of items

  The number of items was $m = 20$ for all studies.

- Number of observations

  The overall sample size was $n = 500$ for all studies.

  Depending on whether DIF or ability differences were simulated, either all responses were generated with the same item and person parameters, or with item and person parameters differing between the groups.

- Item parameters

  The item difficulty parameters were arbitrarily chosen to be:

  $\beta^\mathsf{T} = (0, -0.5, 0, -0.5, -1, -2, -3, -2, -1, 0, +1, +2, +3, +2, +1, +2, +1, 0, -1, 0)$.

  This choice of item parameters was intended to ensure an adequate overlap between the item and person parameter distributions, as illustrated for one sample from a $N(0, 1)$ person parameter distribution in the person-item-map in Figure 4.

  When DIF was simulated, these were the item parameters for the reference group. For the focal group, the value δ was added to the third item parameter.

  The size of δ was varied between 0 and 1.5 in all simulation studies. This choice of δ was made to ensure well comparable results for the LR test and Rasch tree in all scenarios.

- Person parameters

The person ability parameters were drawn from a standard normal distribution $N(0, 1)$ when no ability difference was simulated.

When an ability difference was simulated, the person parameter distributions for the reference and focal groups were simulated with a difference of Δ between their means by drawing the person ability parameters from $N(0 - \frac{\Delta}{2}, 1)$ and $N(0 + \frac{\Delta}{2}, 1)$ respectively.

Negative values of Δ correspond to a scenario where both the DIF and the ability difference disadvantage the focal group, whereas positive values of Δ correspond to a scenario where the DIF disadvantages the focal group while the ability difference favors it.

The ability difference Δ was varied between -0.5 and 0.5 in simulation experiment II and fixed to 0 in all other experiments.

The choice of 0.5 for Δ was intended to ensure that there remains an adequate overlap between the item and person parameter distributions after shifting the person parameter distributions.

3.3. Simulation study I

Rationale of simulation study I

The aim of simulation study I is to illustrate the performance of the Rasch tree and the LR test under the null hypothesis of no DIF and the alternative of DIF being present.

An important aspect of this comparison is the difference between LR test and Rasch tree when handling numeric covariates: For the LR test, reference and focal groups need to be pre-specified. Usually numeric covariates are split at the median to define the two groups. This approach was chosen here for the LR test.

As opposed to that, the Rasch tree has to search over all possible binary partitions of the numeric covariate. This is a disadvantage when compared to the LR test for two given groups when the groups are correctly specified, but may be an advantage when the correct specification is not available.

Moreover, the results of this simulation study would show any inflating effect that the exhaustive search over all possible cutpoints may have on the type I error rate of the method.

Design of simulation study I

The following experimental factors were varied in this simulation experiment:

- Effect size of DIF

  $\delta = 0$ corresponds to the null hypothesis scenario with no DIF, where all responses are generated with the same item parameters. In this scenario, the percentage of significant test results reflects the type I error rate of the method.

  $\delta = 1.5$ corresponds to the alternative scenario with DIF, where the third item is more difficult for the focal group. In this scenario, the percentage of significant test results reflects the statistical power of the method.

- Predictor type

  The type of predictor variable that defines the reference and focal groups was either binary or numeric. The binary predictor variable was sampled from a Binomial distribution with equal class probabilities. The numeric predictor variable was sampled from a discrete uniform distribution over the values 0 to 100.

  When the predictor type was binary, DIF was simulated between the two groups corresponding directly to the two categories of the binary covariate. When the predictor type was numeric, DIF was simulated between two groups specified by values up to the cutpoint and above the cutpoint (the choices of which are described below) in the numeric covariate.

  In this simulation experiment only one covariate – either the binary or the numeric one – was provided to the Rasch tree and LR test (whereas in simulation study III both covariates are provided so that the selection of the correct splitting variable becomes an additional part of the task).

- Cutpoint location

  For the binary covariate, there is only one possible cutpoint by definition. For the numeric covariate, the reference and focal groups were created either by splitting at the median (situated around the value 50) or at the value 80. This variation was chosen to mimic a pattern of DIF in a numeric variable like age, where the difficulty of an item is higher for subjects over a certain age – but not necessarily the median age.

- Test specification

  For the binary covariate, the groups to be tested for DIF correspond directly to the two levels of the covariate and were thus directly provided to both the LR test and the Rasch tree. For the numeric covariate, the LR test was specified such that the groups to be tested for DIF were defined by a median split, whereas the Rasch tree had to search for the optimal cutpoint.

  Note that in this simulation scenario the LR test has an advantage when the true cutpoint is the median, because in these scenarios it was provided with the correct partition, whereas the Rasch tree may have an advantage when the true cutpoint is not the median (further scenarios where the data-driven cutpoint selection of the Rasch tree may be an advantage are investigated in simulation study III).

Results of simulation study I

- Type I error and power

  As can be seen in Table 3, both the LR test and the Rasch tree roughly hold the specified 5% α-level under the null hypothesis of no DIF. Instead of an artificial inflation of the type I error rate due to the exhaustive search for the optimal cutpoint, the Rasch tree even behaves conservatively in the two scenarios involving the numeric covariate. (This is a known behavior of the underlying maximum LM test, where for moderate sample sizes the discrete empirical process cannot fluctuate as much as its continuous asymptotic counterpart, the Brownian bridge, cf. Section 2.2. Besides the sample size,

Method	Predictor	Cutpoint	Type I error	CPU time		
				med	mean	max
LR test	binary		0.054	2.94	2.98	5.27
	numeric	median	0.055	2.94	2.99	5.25
		80	0.054	2.22	2.26	3.34
Rasch tree	binary		0.049	0.68	0.74	2.70
	numeric	median	0.038	0.72	3.22	146.22
		80	0.035	0.48	1.91	72.54

Table 3: Results of simulation study I – no DIF.

Method	Predictor	Cutpoint	Power	RMSE	ARI	CPU time		
						med	mean	max
LR test	binary		0.998	0.211	0.998	2.66	2.71	4.32
	numeric	median	0.999	0.213	0.997	2.72	2.76	4.52
		80	0.282	0.594	0.047	2.28	2.32	3.51
Rasch tree	binary		0.998	0.212	0.998	1.88	1.91	3.38
	numeric	median	0.979	0.326	0.883	43.18	44.46	103.03
		80	0.751	0.410	0.650	41.25	33.13	151.08

Table 4: Results of simulation study I – DIF.

the asymptotic behavior of the maximum LM statistic is also affected by the number of potential cutpoints. For a more detailed discussion of this issue – and how to address it if the number of potential cutpoints is very low, e.g., for numeric variables measured on a coarse grid, but also for ordinal partitioning variables – see Hothorn and Zeileis (2008) and Merkle, Fan, and Zeileis (2013).)

Under the alternative hypothesis of DIF, Table 4 shows that both the LR test and the Rasch tree have a very high power when the reference and focal groups correspond to the two categories of a binary covariate. When the reference and focal groups correspond to the values up to the median and above the median of a numeric covariate (which directly corresponds to the specification of the LR test, whereas the Rasch tree needs to search for the optimal cutpoint), the power of the LR test is somewhat higher than that of the Rasch tree. However, when the reference and focal groups correspond to values up to the value 80 and above the value 80 (i.e. when the LR test is misspecified, whereas the Rasch tree again searches for the optimal cutpoint), the power of the LR test (0.282) is much lower than for the Rasch tree (0.751). This indicates that the Rasch tree is well able to identify the DIF, whereas the LR test is misled by the wrong specification of the median split.

- RMSE and ARI

  (Note that for the RMSE low values are good because they indicate that the simulated item parameters were well recovered, whereas for the ARI high values are good because they indicate a high overlap between the simulated and detected groups.)

  Table 4 shows that under the alternative hypothesis of DIF in the scenario where the reference and focal groups correspond to the values up to the median and above the median of a numeric covariate (which again directly corresponds to the specification of

Method	Cutpoint	Avg. estimated Cutpoint	Bias	Variance	MSE
LR test	median	50.02	0.00	5.01	5.01
	80	50.04	-29.96	5.10	902.42
Rasch tree	median	49.94	-0.08	33.09	33.10
	80	78.42	-1.58	83.03	85.52

Table 5: Results of simulation study I – Cutpoint estimation in numeric predictor.

the LR test, whereas the Rasch tree needs to search for the optimal cutpoint), the RMSE of the LR test (0.213) is lower than that of the Rasch tree (0.326) and the ARI of the LR test (0.997) is slightly higher than that of the Rasch tree (0.883). The reason for this is that the LR test is provided with the correct specification of the groups corresponding to the two different item parameter values, whereas the Rasch tree has to search for the correct groups and in some replications will miss them.

However, in the scenario where the reference and focal groups correspond to values up to the value 80 and above the value 80 (i.e. when the LR test is misspecified, whereas the Rasch tree again searches for the optimal cutpoint), the RMSE of the Rasch tree (0.41) is lower than that of the LR test (0.594) and the ARI of the Rasch tree (0.65) is much higher than that of the LR test (0.047). This indicates that the Rasch tree is still able to identify the correct groups corresponding to the two different item parameter values in many cases, whereas the LR test is misled by the wrong specification of the median split.

- Cutpoint estimation

  Table 5 gives a more detailed analysis of the cutpoint estimation, that is a crucial part of correctly recovering reference and focal group. For the LR test, where the median is always specified as the cutpoint in the numeric variable, the average estimated cutpoint is close to the expected median 50. For the Rasch tree, on the other hand, the average estimated cutpoint reflects the actual simulation design with an average estimated cutpoint close to 50 when the true cutpoint is the median and an average estimated cutpoint close to 80 when the true cutpoint is 80.

  Accordingly, the LR test produces an unbiased estimate by definition when the true cutpoint is the median, but shows a systematic bias when the true cutpoint is not the median. Due to its data-driven approach, the Rasch tree produces virtually unbiased estimates in both scenarios. (The slight negative bias in the second scenario is only due to the fact that the true cutpoint 80 is rather close to the upper boundary of the partitioning variable, that ranges from 1 to 100. Hence, the sampling distribution of the cutpoint estimator is somewhat left-skewed so that its mean is slightly lower than 80, whereas its median is exactly 80.)

  The variance of the estimation is larger for the Rasch tree than for the LR test, as expected (for the LR test, the variance over the iterations merely reflects the variance of the sample median whereas for the Rasch tree, the variance actually includes the variance of the data-driven cutpoint estimation). However, when considering the combination of squared bias and variance in the MSE, the systematic bias of the LR test in

the scenario where the true cutpoint is not the median results in the – by far – highest MSE.

- Computation time

  For those scenarios with the binary covariate the computation times for the Rasch tree are substantially lower than for the LR test in Tables 3 and 4, as is to be expected from the test construction principles outlined in Section 2.3. On the other hand, in those scenarios where the Rasch tree needs to search over all possible cutpoints in the numeric variable, the maximum (and to some extent the mean) computation times are already increased under the null hypothesis of no DIF, as displayed in Table 3, whereas the median computation times are notably increased only under the alternative hypothesis of DIF, as displayed in Table 4. This is to be expected from the construction of the Rasch tree method, where the computationally expensive selection of the optimal cutpoint is only conducted in those cases where significant parameter instability is detected in the first place, as outlined in Sections 2.3 and 2.4. Therefore, the computation times for the Rasch tree method are high only in those scenarios where significant parameter instability is present and the tree needs to search over all possible cutpoints in the numeric covariate, whereas the computational effort is very low in all other scenarios.

Conclusions from simulation study I

From the results of simulation study I it becomes clear that – despite the exhaustive search for the optimal cutpoint in numeric variables – the Rasch tree approach is not affected by an inflation of the type I error rate under the null hypothesis. Under the alternative hypothesis, where DIF is present, it shows a power comparable to that of the LR test when the correct partition is known and provided to the LR test, but a notably higher power (and also a better recovery of the simulated reference and focal groups, as indicated by RMSE, ARI and the quality of the cutpoint estimation) when the correct partition is not known. For practical applications, this means that the LR test may miss DIF in a numeric variable due to the wrong – yet very common – specification based on the median split, whereas the Rasch tree approach has a much higher chance of detecting DIF in this situation.

3.4. Simulation study II

Rationale of simulation study II

The aim of simulation study II is to illustrate the effect of a true ability difference between reference and focal group on the type I error rate and power of the LR test and Rasch tree. In particular, a test for DIF should not be misled towards an inflated type I error rate by an ability difference between reference and focal group when no DIF is present.

Design of simulation study II

The following experimental factors were varied in this simulation experiment:

- Effect size of DIF

Method	Ability difference	Type I error	CPU time		
			med	mean	max
LR test	−0.5	0.059	1.84	1.86	2.65
	+0.5	0.055	1.88	1.90	2.58
Rasch tree	−0.5	0.051	0.56	0.61	1.93
	+0.5	0.043	0.58	0.61	1.96

Table 6: Results of simulation study II – no DIF.

$\delta = 0$ again corresponds to the null hypothesis scenario with no DIF, where all responses are generated with the same item parameters. In this scenario, the percentage of significant test results reflects the type I error rate of the method.

$\delta = 1.5$ again corresponds to the alternative scenario with DIF, where the third item is more difficult for the focal group. In this scenario, the percentage of significant test results reflects the statistical power of the method.

- Ability difference

  As opposed to simulation study I, where there was no ability difference between the groups, now the reference and focal groups also differ in their mean abilities by the value $\Delta = -0.5$ or 0.5.

A binary predictor variable, sampled again from a Binomial distribution with equal class probabilities, was used in all scenarios of simulation study II.

Results of simulation study II

- Type I error and power

  As can be seen in Table 6, both the LR test and the Rasch tree roughly hold the specified 5% α-level under the null hypothesis of no DIF. Instead of an artificial inflation of the type I error rate due to the ability difference between reference and focal group, the Rasch tree even behaves somewhat conservatively in the scenario where the ability difference $\Delta = 0.5$ is in favor of the focal group. (The reason for this seems to be that with $\Delta = 0.5$ the extreme item parameters were not well enough covered by the person distributions in both groups, so that their estimates show a higher variance. The resulting hetereoskedasticity in the cumulative sum process apparently renders the LM test employed in the Rasch tree, whose asymptotics do not take effect as quickly as for the LR test, somewhat conservative.)

  Under the alternative hypothesis, where both an ability difference and DIF are present, Table 7 shows that both the LR test and the Rasch tree have a very high power to detect the DIF.

- Computation time

  The computation times again show an advantage of the Rasch tree in the presence of the binary covariate, especially in Table 6 under the null hypothesis.

Method	Ability difference	Power	CPU time		
			med	mean	max
LR test	−0.5	0.997	2.52	2.52	3.61
	+0.5	0.998	2.41	2.41	3.08
Rasch tree	−0.5	0.997	1.62	1.61	2.17
	+0.5	0.998	1.63	1.62	2.33

Table 7: Results of simulation study II – DIF.

Conclusions from simulation study II

Both the LR test and the Rasch tree approach are not misled towards an increased type I error rate in the presence of ability differences and still show a high power for detecting DIF in the alternative scenarios of this simulation study, where the correct partition was provided by the binary predictor variable.

3.5. Simulation study III

Rationale of simulation study III

Whereas the LR test can only detect DIF in previously specified groups, the Rasch tree searches over all provided covariates and all possible cutpoints. Therefore, the aim of simulation study III is to illustrate how LR test and Rasch tree perform in the presence of focal and reference groups that result from non-standard patterns, such as non-median splits, u-shaped patterns and interactions of covariates – none of which would typically be specified in a LR test.

Since in this simulation study both covariates are presented at a time, another aspect of interest is a potential inflating effect that the multiple testing over more than one covariate may have on the type I error rate of the methods.

Design of simulation study III

The following experimental factors were varied in this simulation experiment:

- Effect size of DIF

  $\delta = 0$ again corresponds to the null hypothesis scenario with no DIF, where all responses are generated with the same item parameters. In this scenario, the percentage of significant test results reflects the type I error rate of the method.

  $\delta = 1.5$ again corresponds to the alternative scenario with DIF, where the third item is more difficult for the focal group. In this scenario, the percentage of significant test results reflects the statistical power of the method.

- Pattern of reference and focal groups

  When the pattern was binary, DIF was simulated between the two groups corresponding directly to the two categories of the binary covariate, like in simulation studies I and II.

When the pattern was numeric, DIF was simulated between two groups specified by a value of the numeric covariate up to and above a certain cutpoint (the choices of which are described below), like in simulation study I.

When the pattern was u-shaped, DIF was simulated between two groups specified by values of the numeric covariate up to the value 20 and from the value 80 vs. values between 20 and 80. This variation was chosen to mimic a pattern of DIF in a numeric variable like age, where DIF is present for young and old subjects as opposed to middle-aged subjects.

When the pattern was interaction, DIF was simulated between two groups specified by those observations with a value of 1 in the binary covariate *and* a value of the numeric covariate above a certain cutpoint (the choices of which are described below) vs. all other observations. This variation was chosen to mimic a pattern of DIF that depends on more than one variable, such as a combination of age and gender.

- Cutpoint location

  In those patterns involving the numeric covariate, the groups were again created either by splitting at the median (situated around the value 50) or at the value 80. This variation was again chosen to mimic a pattern of DIF in a numeric variable like age, where the difficulty of an item is higher for subjects over a certain age – but not necessarily the median age.

- Test specification

  For the LR test, again two typical specifications were made, where either the two groups to be tested corresponded directly to the two levels of the binary covariate or the two groups were specified by a median split in the numeric covariate. These specifications coincide with the binary pattern and the numeric pattern with median split, whereas they can be considered as misspecifications for the numeric pattern with a split in another cutpoint as well as for the u-shaped and the interaction patterns.

  For the Rasch tree no specification is necessary.

The binary and numeric covariates were sampled from the same distributions as described before. Moreover, in this experiment – as opposed to the previous ones – both covariates were provided to the methods in each replication (one at at time to the LR test and both simultaneously to the Rasch tree).

Under these circumstances the results for the type I error rate and power of the two methods would not be directly comparable if the Rasch trees were computed with the suggested Bonferroni adjustment but the LR tests were not, because both methods are provided with both covariates and are thus equally affected by multiple testing. Therefore, we have implemented a Bonferroni adjustment for the p values resulting from the LR test, too, and display all results with and without Bonferroni adjustment for both methods.

Note also that the fact that both covariates are provided in each replication means that for the Rasch tree the selection of the correct splitting variable(s) is now part of the task, whereas for the LR test, where a pre-specification of the reference and focal groups is necessary, it means that in some scenarios the test was computed for the "wrong" variable (which corresponds to a null hypothesis scenario, so that the reported power actually reflects the type I error rate in these scenarios).

Method	Pattern	Cutpoint	Specification	Type I error with Bonf. adj.	Type I error without Bonf. adj.
LR test	binary		binary	0.027	0.054
			numeric	0.027	0.053
	numeric	median	binary	0.026	0.049
			numeric	0.025	0.047
		80	binary	0.031	0.054
			numeric	0.026	0.047
	u-shaped		binary	0.030	0.053
			numeric	0.025	0.052
	interaction	median	binary	0.028	0.057
			numeric	0.031	0.055
		80	binary	0.025	0.052
			numeric	0.025	0.051
Rasch tree	binary			0.038	0.080
	numeric	median		0.040	0.082
		80		0.043	0.088
	u-shaped			0.040	0.079
	interaction	median		0.039	0.088
		80		0.037	0.083

Table 8: Results of simulation study III – no DIF.

The last point that should be noted for this experiment is that the power is no longer the ideal measure to compare the performance of the two methods, especially in those scenarios where the correct pattern of reference and focal groups can only be replicated by means of more than one split in one or both covariates, because the way the power is computed for both methods makes it hard to compare them directly in these complex scenarios: For the LR test the power is computed as the percentage of replications in which a test for DIF for the two pre-specified groups returned a significant result. For the Rasch tree, however, the power is computed as the percentage of replications in which at least one split is made by the tree – which indicates whether any DIF is detected at all, but does not reflect the fact that the tree actually provides much more information about the group pattern. Therefore, in this experiment it is particularly helpful to consider not only the results for the power but also for RMSE and the ARI, that express how well the simulated group pattern was recovered.

To save space, the computation times will not be presented for this simulation study because they are no longer comparable due to the fact that the Rasch tree had to process both predictor variables simultaneously in each scenario, whereas the LR test only processed one variable at a time.

Results of simulation study III

- Type I error

  As can be seen in Table 8 for the type I error rates resulting from the Bonferroni adjusted

				with Bonf. adj.			without Bonf. adj.		
Method	Pattern	Cutpoint	Specification	Power	RMSE	ARI	Power	RMSE	ARI
LR test	binary		binary	0.996	0.214	0.996	0.998	0.211	0.998
			numeric	0.026	0.765	0.000	0.052	0.766	0.000
	numeric	median	binary	0.025	0.766	0.000	0.050	0.766	0.000
			numeric	0.996	0.217	0.993	0.999	0.214	0.996
		80	binary	0.029	0.611	0.000	0.052	0.612	0.000
			numeric	0.189	0.600	0.032	0.280	0.595	0.047
	u-shaped		binary	0.027	0.753	0.000	0.051	0.754	0.000
			numeric	0.026	0.754	0.000	0.050	0.754	0.000
	interaction	median	binary	0.437	0.628	0.108	0.556	0.616	0.137
			numeric	0.439	0.627	0.114	0.551	0.615	0.142
		80	binary	0.063	0.472	0.003	0.105	0.473	0.004
			numeric	0.058	0.471	0.003	0.109	0.472	0.005
Rasch tree	binary			0.995	0.222	0.985	0.998	0.219	0.988
	numeric	median		0.964	0.339	0.864	0.980	0.333	0.862
		80		0.658	0.443	0.560	0.762	0.417	0.624
	u-shaped			0.515	0.644	0.277	0.651	0.607	0.354
	interaction	median		0.514	0.579	0.188	0.650	0.552	0.238
		80		0.169	0.469	0.053	0.265	0.466	0.078

Table 9: Results of simulation study III – DIF.

p values (left column for type I error), the Rasch tree does not exceed the specified 5% α-level but again behaves slightly conservatively when both variables are presented at a time. The corresponding type I error rates for the LR test have to be added for the binary and the numeric splitting variable in each scenario for comparison, in which case the LR test roughly holds the specified 5% α-level when taking both variables together.

If no Bonferroni adjustment is applied, the type I error rates in Table 8 (right column for type I error) indicate that the Rasch tree shows an exceeded type I error rate of around 8% when both variables are presented at the same time. The corresponding results for the LR test show an exceeded type I error rate of about 10% when taking both variables together.

The results support the widely known fact that Bonferroni adjustment is rather conservative (and it may be worth considering other options for the Rasch tree), but that some type of adjustment is necessary for any method when more than one covariate is investigated for DIF at the same time.

- Power, RMSE and ARI

  As can be seen in Table 9, the power of the LR test depends strongly on the correspondence between the simulated scenario and the specification of the test:

  When the reference and focal groups correspond directly to the two categories of the binary variable and the binary variable is used for specifying the groups in the LR test, or when the reference and focal groups correspond to values up to the median and above the median of the numeric variable and a median split is used for specifying the groups in the LR test, the power is very high, just like in simulation study I.

If, on the other hand, the wrong variable is provided to the LR test, like in those scenarios where the simulated pattern is binary but the variable provided to the LR test is the numeric one and vice versa, its power reflects the type I error rate under the null hypothesis (of about 2.5% with Bonferroni adjustment, left column for the power, and about 5% without Bonferroni adjustment, right column for the power in Table 9), as is to be expected.

The more interesting results correspond to those scenarios where focal and reference groups are defined by non-median splits, u-shaped patterns and combinations of covariates. (For readability, we will only refer to the results for the unadjusted p values from the right columns in Table 9 in the following, because the unadjusted results are directly comparable to those of simulation study I, but of course the Bonferroni adjusted results show the same pattern):

When the reference and focal groups correspond to a split in the numeric variable that is not located at the median, the LR test has a much lower power than the Rasch tree (0.280 vs. 0.762), just like in simulation study I (where any numerical differences in the criterion variables are only due to random variation).

The disadvantage of the LR test is even more pronounced when the reference and focal groups are defined by a u-shaped pattern in the numeric variable (0.050 vs. 0.651), in which case the power of the LR test is at the same level as under the null hypothesis while the Rasch tree is still able to detect the DIF in many cases.

For the interaction pattern in the scenario where the interaction is formed with a median split in the numeric variable, the LR test still has a rather large power, no matter whether it uses the binary or the numeric variable for defining the groups (0.556 using only the binary and 0.551 using only the numeric variable for defining the groups), because either split creates one pure group and one group for which about half of the observations have been generated with a different item parameter, so that the LR test still has a good chance to detect the DIF based on either variable alone. Yet the power of the Rasch tree in this scenario is notably higher (0.650).

While the power only reflects whether any DIF was detected at all, however, the corresponding RMSE (where lower values indicate better recovery of the simulated item parameters) and ARI (where higher values indicate better recovery of the simulated groups) indicate more clearly whether the correct group pattern was recovered:

The RMSE for the Rasch tree (0.552) is notably lower than for the LR test (0.616 using only the binary and 0.615 using only the numeric variable for defining the groups). The ARI, on the other hand, is higher for the Rasch tree (0.238) than for the LR test (0.137 using only the binary and 0.142 using only the numeric variable for defining the groups).

For the most complicated interaction pattern, where the interaction is formed with a split at the value 80 of the numeric variable, the results in Table 9 show that the power of the LR test is now even lower (0.105 using only the binary and 0.109 using only the numeric variable for defining the groups). The power of the Rasch tree in this most complicated scenario is also low (0.265), but more than twice the size of that of the LR test. (The reason for the low overall power in this scenario is that the number of observations in the focal group is now only about 50, as compared to about 125 in the median split scenario. This pattern is harder to detect for both methods but again

particularly hard for the LR test that cannot search for the optimal cutpoint like the Rasch tree, but is restricted to the arbitrarily pre-specified median split.)

The RMSE for the Rasch tree (0.466) is again a little lower than for the LR test (0.473 using only the binary and 0.472 using only the numeric variable for defining the groups). The ARI, on the other hand, is notably higher for the Rasch tree (0.078) than for the LR test (0.004 using only the binary and 0.005 using only the numeric variable for defining the groups).

Note that the reported ARI values actually underestimate the group recovery of the Rasch tree in all settings where the two simulated groups are described by three final nodes in the recursive partitioning structure of the Rasch tree (i.e., in the u-shaped and interaction scenarios), because the ARI cannot reach its theoretical maximum of 1 when the number of simulated and recovered groups is not equal. Yet, the ARI values for the Rasch tree are still notably higher than those for the LR test, especially in the u-shaped scenario where the LR test completely fails to detect the group difference.

Conclusions from simulation study III

Due to the suggested Bonferroni adjustment of the p values, the type I error rates for the Rasch tree method are not inflated even when more than one covariate is presented at the same time. Since multiple testing affects any test for DIF when more than one covariate is investigated at the same time, it should be noted that some type of α-adjustment is necessary for any DIF detection method in this case.

With respect to power, simulation study III has shown that in all scenarios where the typical specification for the LR test does not correspond to the actual (but in reality unknown) pattern of DIF present in the data, the more flexible Rasch tree approach clearly outperforms the standard LR test – both with respect to the power for detecting DIF in the first place and with respect to correctly recovering the groups with different item parameters, which is of high interest in practical applications.

To further illustrate the practical importance of this finding, we will now show in a small application example that a non-standard pattern of DIF – such as an interaction of two variables – is interesting not only from a theoretical point of view for simulation studies, but is also a realistic scenario for empirical data.

4. Application example

An online quiz for testing one's general knowledge was conducted by the weekly German news magazine SPIEGEL in 2009. Overall, about 700,000 respondents participated in the quiz and answered a set of sociodemographic questions. The general knowledge quiz consisted of a total of 45 items from five different domains: politics, history, economy, culture, and natural sciences. For each domain, four different sets of nine items were available, that were randomly assigned to the participants. A thorough discussion and analysis of the original data set is provided in Trepte and Verbeet (2010).

For further illustration of the Rasch tree method, we consider only an exemplary selection of subjects, covariates and items: To limit the number of participants to a sample size realistic for psychological research, we included only university students enrolled in the federal state

Variable	Summary statistics					
Gender	male: 415			female: 641		
	$x_{\min}$	$x_{0.25}$	x_{med}	$\bar{x}$	$x_{0.75}$	$x_{\max}$
Age	18	21	23	23.09	25	40

Table 10: Summary statistics for the considered covariates.

of Bavaria, who had been assigned questionnaire number 20. This sample still contains 1056 complete cases, that were employed in the following analysis. To avoid any obvious multidimensionality (the data were originally analyzed as if they were unidimensional, but from the construction of the quiz it appears that the different domains should be treated as separate dimensions), we also limited our consideration to only one domain – history – with nine items. To test for DIF in this supposedly unidimensional scale, we employed the covariates gender and age, whose summary statistics are provided in Table 10.

The nine items included in the history knowledge scale (with the correct answers) were:

1. The Roman naval supremacy was established through. . . – . . . the abolition of Carthage.
2. In which century did the Thirty Years' War take place? – The 17th century.
3. Which form of government is associated with the French King Louis XIV? – Absolutism.
4. What island did Napoleon die on in exile? – St. Helena.
5. How many percent of the votes did the NSDAP receive in the 1928 elections of the German Reichstag? – About 3 percent.
6. How many Jews were killed by the Nazis during the Holocaust? – About 6 Million.
7. Who is this? – (Picture of Johannes Rau, former German federal president.)
8. Which of the following countries is not a member of the EU? – Croatia[1].
9. How did Mao Zedong expand his power in China? – The Long March.

The Rasch tree resulting for this exemplary data set is depicted in Figure 5. The computation time for this analysis was 1.768 seconds on the server and 1.064 seconds on a laptop with an Intel Core processor with 2.53GHz.

The mere fact that the Rasch tree displays more than one terminal node means that measurement invariance cannot be assumed and the history knowledge of the participants should not be compared by means of one joint Rasch model. In particular, Figure 5 shows that DIF is present between females, males up to the age of 22, and males above the age of 22. The Rasch tree result thus illustrates that it is an interaction of gender and age that determines the groups exhibiting DIF in this exemplary data set.

With standard approaches, this pattern could only be detected if the interaction term was explicitly included in the model or the respective groups (including the correct cutpoint in the numeric variable) were explicitly provided in the specification of the test. However, in practice usually only DIF in single variables is investigated, so that an interaction structure like in this example would not be detected.

[1]At the time the quiz was conducted, Croatia was not yet a member of the EU.

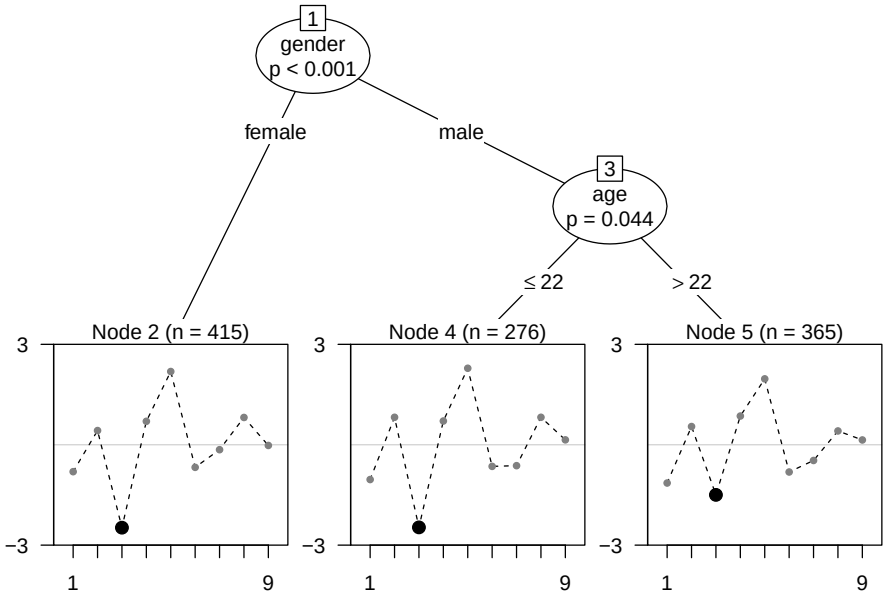

Figure 5: Rasch tree for the general knowledge quiz example.

With respect to the items affected by the DIF, it is most prominent from Figure 5 that, relative to the other items, the third item (that was manually highlighted here for illustration purposes) is easier for females and younger males as compared to older males. Taking this item as an example, the content of the item (it refers to the form of government associated with the French King Louis XIV) together with the information which subjects find the item easier or harder to solve (it is easier for females and younger males) can help content experts generate hypotheses about possible sources of the DIF (such as a higher exposure to the figure or era of Louis XIV in history or french class – or in recent TV shows).

5. Discussion and outlook

We have proposed a new method for detecting DIF in the Rasch model, that combines the advantages of previous approaches for given groups and latent classes: Groups of subjects exhibiting DIF are automatically detected, but remain directly interpretable with respect to their covariate values. In particular, in numeric covariates it is no longer necessary to pre-specify a cutpoint for defining focal and reference groups, but the cutpoint associated with the strongest parameter difference is detected automatically. Thus, DIF in a numeric covariate cannot go unnoticed due to a suboptimal definition of the groups.

Of course, any covariate-based approach can only detect all groups of subjects with DIF when all relevant covariates are available for the analysis. If a covariate causing DIF has been missed in the data aquisition, the algorithm has no chance to detect it (yet the DIF may be detected if another covariate, that is correlated with the missing one, is available for splitting). Moreover, just like different combinations of covariates can yield the same prediction in regression models, different combinations of covariates may also predict the parameter profiles similarly well as the ones selected by a Rasch tree for a given sample.

Therefore it should be noted that – as with all observational data – a covariate used for splitting cannot simply be interpreted as the causal source of the observed DIF, because the observed splitting variable may only serve as a proxy for the unobserved (and potentially unobservable) true cause. For example, if DIF is detected between men and women, gender should not be considered as the actual cause of the DIF, but as an indicator of a variety of educational and social influences.

Beside the use of person covariates to detect groups of subjects with different item parameters, the method suggested here is also applicable for detecting DIF over time (in the sense of item drift). For this purpose, the variable time (measured numerically or at two or more discrete time points) can be employed for splitting just like any other covariate. As a result, one or more splits in the Rasch tree would indicate an instability in the item parameters over time, which would need to be accounted for in longitudinal comparisons.

In order to help applied researchers with the interpretation of DIF for specific items and specific groups, in the future we will try to provide additional means of visualization and post-hoc item-wise comparisons after the global DIF tests conducted by the Rasch tree. Moreover, we are currently working on generalizations of the Rasch tree method to extensions of the Rasch model. In particular, a generalization of the Rasch tree method to the partial credit model (Masters 1982) will be used to detect both differential item and differential step functioning (Penfield 2007; Penfield, Alvarez, and Lee 2009). Other interesting extensions that we will try to address in future work are the generalization to the 2PL or Birnbaum model (Birnbaum 1968), that may prove helpful for the analysis of nonuniform DIF, and the generalization to a 2-parameter logistic model including a location and a guessing parameter, because this would allow the detection of differential guessing behavior in the case of multiple choice items (also investigated by Ben-Shakhar and Sinai 1991 and Westers and Kelderman 1992). A related method for detecting different preferences between groups of subjects in the Bradley-Terry model (Strobl, Wickelmaier, and Zeileis 2011) is already implemented in the `psychotree` package.

Computational details

Our results were obtained using the R system for statistical computing (R Development Core Team 2012), version 3.2.0, and the add-on package `psychotree` (Zeileis *et al.* 2012), version 0.12-3.

For the simulation studies, we also employed functions from the add-on packages `eRm` 0.15-4. (Mair and Hatzinger 2007; Mair *et al.* 2012) and `ltm` 1.0-0 (Rizopoulos 2006, 2012). The person-item-map in Figure 4 was drawn by means of the `plotPImap` function available in the `eRm` package. Package `mclust` 4.3 (Fraley and Raftery 2002, 2012) was utilized for computing the adjusted Rand index.

All packages are freely available under the General Public License from the Comprehensive R Archive Network. A vignette describing the practical application of the Rasch tree method is available along with the `psychotree` package at `http://CRAN.R-project.org/package=psychotree/`.

Acknowledgments

Julia Kopf is supported by the German Federal Ministry of Education and Research (BMBF) within the project "Heterogeneity in IRT-Models" (grant ID 01JG1060).

The authors would like to thank three anonymous reviewers for their very helpful and constructive feedback.

Special thanks go to Reinhold Hatzinger (1953–2012), who has stimulated important insights for this and other projects through many conversations and his extensive work on the R package `eRm` (Mair and Hatzinger 2007; Mair *et al.* 2012). We miss him as a researcher and friend.

References

Andersen E (1972). "A Goodness of Fit Test for the Rasch Model." *Psychometrika*, **38**, 123–140.

Andrews DWK (1993). "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica*, **61**, 821–856.

Ben-Shakhar G, Sinai Y (1991). "Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies." *Journal of Educational Measurement*, **28**(1), 23–35.

Birnbaum A (1968). "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In F Lord, M Novick (eds.), *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading.

Boulesteix AL (2006). "Maximally Selected Chi-Square Statistics and Binary Splits of Nominal Variables." *Biometrical Journal*, **48**(5), 838–848.

Breiman L, Friedman J, Olshen R, Stone C (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Cohen A, Bolt D (2005). "A Mixture Model Analysis of Differential Item Functioning." *Journal of Educational Measurement*, **42**(3), 133–148.

Dobra A, Gehrke J (2001). "Bias Correction in Classification Tree Construction." In CE Brodley, AP Danyluk (eds.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA*, pp. 90–97. Morgan Kaufmann.

Fischer G, Molenaar I (eds.) (1995). *Rasch Models: Foundations, Recent Developments and Applications*. Springer-Verlag, New York.

Fraley C, Raftery A (2002). "Model-Based Clustering, Discriminant Analysis and Density Estimation." *Journal of the American Statistical Association*, **97**(458), 611–631.

Fraley C, Raftery A (2012). `mclust`*: Model-Based Clustering/Normal Mixture Modeling*. R package version 3.4.11, URL `http://CRAN.R-project.org/package=mclust`.

Gelin M, Carleton B, Smith M, Zumbo B (2004). "The Dimensionality and Gender Differential Item Functioning of the Mini Asthma Quality of Life Questionnaire (MiniAQLQ)." *Social Indicators Research*, **68**, 91–105.

Gustafsson J (1980). "Testing and Obtaining Fit of Data in the Rasch Model." *British Journal of Mathematical and Statistical Psychology*, **33**(2), 205–233.

Hancock G, Samuelsen K (eds.) (2007). *Advances in Latent Variable Mixture Models*. Information Age Publishing, Charlotte.

Hochberg Y, Tamhane A (eds.) (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

Hothorn T, Lausen B (2003). "On the Exact Distribution of Maximally Selected Rank Statistics." *Computational Statistics & Data Analysis*, **43**(2), 121–137.

Hothorn T, Zeileis A (2008). "Generalized Maximally Selected Statistics." *Biometrics*, **64**(4), 1263–1269.

Hubert L, Arabie P (1985). "Comparing Partitions." *Journal of Classification*, **2**(1), 193–218.

Kelderman H, MacReady G (1990). "The Use of Loglinear Models for Assessing Differential Item Functioning across Manifest and Latent Examinee Groups." *Journal of Educational Measurement*, **27**(4), 307–327.

Koziol J (1991). "On Maximally Selected Chi-Square Statistics." *Biometrics*, **47**(4), 1557–1561.

Liou M (1994). "More on the Computation of Higher-Order Derivatives on the Elementary Symmetric Functions in the Rasch Model." *Applied Psychological Measurement*, **18**(1), 53–62.

Maij-de Meij A, Kelderman H, Van der Flier H (2008). "Fitting a Mixture Item Response Theory Model to Personality Questionnaire Data: Characterizing Latent Classes and Investigating Possibilities for Improving Prediction." *Applied Psychological Measurement*, **32**(8), 611–631.

Mair P, Hatzinger R (2007). "Extended Rasch Modeling: The `eRm` Package for the Application of IRT Models in R." *Journal of Statistical Software*, **20**(9), 1–20. URL http://www.jstatsoft.org/v20/i09/.

Mair P, Hatzinger R, Maier M (2012). `eRm`*: Extended Rasch Modeling*. R package version 0.15-0, URL http://CRAN.R-project.org/package=eRm.

Marcus R, Peritz E, Gabriel K (1976). "Closed Testing Procedures with Special Reference to Ordered Analysis of Variance." *Biometrika*, **63**(3), 655–660.

Masters G (1982). "A Rasch Model for Partial Credit Scoring." *Psychometrika*, **47**(2), 149–174.

Merkle EC, Fan J, Zeileis A (2013). "Testing for Measurement Invariance with Respect to an Ordinal Variable." *Psychometrika*. (Forthcoming).

Merkle EC, Zeileis A (2013). "Tests of Measurement Invariance without Subgroups: A Generalization of Classical Methods." *Psychometrika*, **78**(1), 59–82.

Miller R, Siegmund D (1982). "Maximally Selected Chi Square Statistics." *Biometrics*, **38**(4), 1011–1016.

Milligan G, Cooper M (1986). "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis." *Multivariate Behavioral Research*, **21**(4), 441–458.

Mislevy R, Verhelst N (1990). "Modeling Item Responses when Different Subjects Employ Different Solution Strategies." *Psychometrika*, **55**(2), 195–215.

Pedraza O, Graff-Radford N, Smith G, Ivnik R, Willis F, Petersen R, Lucas J (2009). "Differential Item Functioning of the Boston Naming Test in Cognitively Normal African American and Caucasian Older Adults." *Journal of the International Neuropsychological Society*, **15**(05), 758–768.

Penfield D (2007). "Assessing Differential Step Functioning in Polytomous Items Using a Common Odds Ratio Estimator." *Journal of Educational Measurement*, **44**(3), 187–210.

Penfield D, Alvarez K, Lee O (2009). "Using a Taxonomy of Differential Step Functioning to Improve the Interpretation of DIF in Polytomous Items: An Illustration." *Applied Measurement in Education*, **22**(1), 61–78.

Perkins A, Stump T, Monahan P, McHorney C (2006). "Assessment of Differential Item Functioning for Demographic Comparisons in the MOS SF-36 Health Survey." *Quality of Life Research*, **15**, 331–348.

R Development Core Team (2012). R*: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P (2003). "A Nonlinear Mixed Model Framework for Item Response Theory." *Psychological Methods*, **8**(2), 185–205.

Rizopoulos D (2006). "`ltm`: An R Package for Latent Variable Modeling and Item Response Analysis." *Journal of Statistical Software*, **17**(5). URL http://www.jstatsoft.org/v17/i05/.

Rizopoulos D (2012). `ltm`*: Latent Trait Models under IRT*. R package version 0.9-9, URL http://CRAN.R-project.org/package=ltm.

Rost J (1990). "Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis." *Applied Psychological Measurement*, **14**(3), 271–282.

Shih YS (2004). "A Note on Split Selection Bias in Classification Trees." *Computational Statistics & Data Analysis*, **45**(3), 457–466.

Smit J, Kelderman H, Van der Flier H (2000). "The Mixed Birnbaum Model: Estimation using Collateral Information." *Methods of Psychological Research Online*, **5**, 1–13.

Strobl C, Boulesteix AL, Augustin T (2007). "Unbiased Split Selection for Classification Trees Based on the Gini Index." *Computational Statistics & Data Analysis*, **52**(1), 483–501.

Strobl C, Malley J, Tutz G (2009). "An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests." *Psychological Methods*, **14**(4), 323–348.

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.

Trepte S, Verbeet M (eds.) (2010). *Allgemeinbildung in Deutschland – Erkenntnisse aus dem SPIEGEL Studentenpisa-Test*. VS Verlag, Wiesbaden.

Van den Noortgate W, De Boeck P (2005). "Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models." *Journal of Educational and Behavioral Statistics*, **30**(4), 443–464.

Westers P, Kelderman H (1992). "Examining Differential Item Functioning due to Item Difficulty and Alternative Attractiveness." *Psychometrika*, **57**(1), 107–118.

Woods C, Oltmanns T, Turkheimer E (2009). "Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality." *Journal of Psychopathology and Behavioral Assessment*, **31**, 320–330.

Zeileis A, Hornik K (2007). "Generalized M-Fluctuation Tests for Parameter Instability." *Statistica Neerlandica*, **61**(4), 488–508.

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

Zeileis A, Strobl C, Wickelmaier F, Kopf J (2012). `psychotree`*: Recursive Partitioning Based on Psychometric Models*. R package version 0.12-2, URL `http://CRAN.R-project.org/package=psychotree`.

Affiliation:

Carolin Strobl
Department of Psychology
Universität Zürich
Binzmühlestr. 14
CH-8050 Zürich, Switzerland
E-mail: `Carolin.Strobl@psychologie.uzh.ch`

Julia Kopf
Department of Statistics
Ludwig-Maximilians-Universität München
Ludwigstr. 33
DE-80539 München, Germany
E-mail: Julia.Kopf@stat.uni-muenchen.de

Achim Zeileis
Department of Statistics
Universität Innsbruck
Universitätsstr. 15
AT-6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org