On the detection of DIF under higher level IRT models using Rasch-trees.

Master's thesis

Dorina Kohler

Student number: 3917970

Reviewer: Prof. Dr. Jürgen Heller

Second reviewer: Ph.D Florian Wickelmaier

Forschungsmethoden und Mathematische Psychologie

Eberhard Karls Universität Tübingen

Statement of authorship:

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst habe, dass ich keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, dass ich alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe, dass die Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist, dass ich die Arbeit weder vollständig noch in wesentlichen Teilen bereits veröffentlicht habe, dass das in Dateiform eingereichte Exemplar mit dem eingereichten gebundenen Exemplar übereinstimmt.

29th October 2020, Tübingen

Abstract

Rasch-trees are a flexible method for detecting differential item functioning (DIF) in data where the Rasch model applies. The current work explores the effect of model misspecification on Rasch-trees, more precisely the performance of Rasch-trees as a global DIF test in data where the 2PL or 4PL model applies. Different conditions were simulated varying the DIF value, the number of items affected by DIF, the parameter(s) that were affected, the underlying model as well as if the covariable, which defines the DIF, was categorical or continuous. DIF was always simulated in the item discrimination for the 2PL model and in the pseudo guessing and slipping error parameter for the 4PL model. In some conditions, DIF was simulated in the item easiness parameter as well. The performance of the Rasch-trees was evaluated in regard to the Power, Type-I-Error rate and the Root mean square error (RMSE) of the item easiness parameter estimation. The method performed better with data under the 2PL model than under the 4PL model. The Power was higher, the RMSE was not augmented unconditionally, the number of identified subgroups was closer to the simulated value. Regarding the 4PL model, DIF in a continuous covariable yielded better results. The Type-I-Error rate was generally not elevated.

Keywords: Rasch-trees, Differential item functioning, Rasch, 2PL, 4PL, Item response theory, Model misspecification

Contents

Since he introduced his famous model in "An individualistic approach to item analysis" (Rasch, 1966), Georg Rasch's name is known to anyone even vaguely interested in test theory. Looking back, one would have to call his work inspiring. Since then and up until today, his model served as the basis of countless articles and works. Over a million articles can be found with the keyword "Rasch" in the title on Google scholar alone (retrieved on the 2nd of October, 2020). The current thesis is included in this list, as is Strobl, Kopf, and Zeileis (2015), which developed a recursive method of detecting differential item functioning (DIF) on the basis of the Rasch model and named it Rasch-trees. The current thesis investigates this method and its performance under model misspecification. For that, certain restrictions set by Rasch himself are overstepped and the consequences are explored. Before that, Item response models, including the Rasch model, are introduced. An explanation of Rasch-trees and the related problem of DIF follows. The motivation, methods, and results of various simulations, which explore the use of Rasch-trees on data, which was not simulated under the Rasch model, are presented and discussed.

Item response theory models

Models of the item response theory (IRT) are centred around the probability of a person's answer to an item. In the case of the dichotomous Rasch model, this probability is given by Formula 1. Polytomous expansion of the Rasch model and other IRT models exist (Andrich, 1978; Masters, 1982; Samejima, 1969). However, the thesis at hand will not consider these due to the base level explorative nature of the work. All models are used in their dichotomous form.

$$P(X_{ij} = 1 \mid \theta_i, b_j) = \frac{e^{\theta_i + b_j}}{1 + e^{\theta_i + b_j}} \tag{1}$$

Formula 1: Probability of item j being answered correctly by person i under the Rasch model with θ_i representing the person's ability and b_j representing the item's easiness.

A helpful visual representation for IRT models are the item characteristic curves (ICCs). Based on Formula 1, an ICC of the Rasch model depicts the probability of a specific item (easiness) dependent on the latent dimension θ. A person's ability θ_i

as well as an item's easiness b_j are located on this spectrum. Figure 1 (left) shows ICCs for three items according to the Rasch model. The easiness b_j of an item can be read off by determining the point when the ICC surpasses $\mathrm{P}(X_j = 1|\, \theta,\, b_j) = .5$ on the continuous latent dimension θ. Note that the curves all have the same slope and only differ in their position on the abscissa. This depicts the specifications of the Rasch model as the most restrained IRT model.

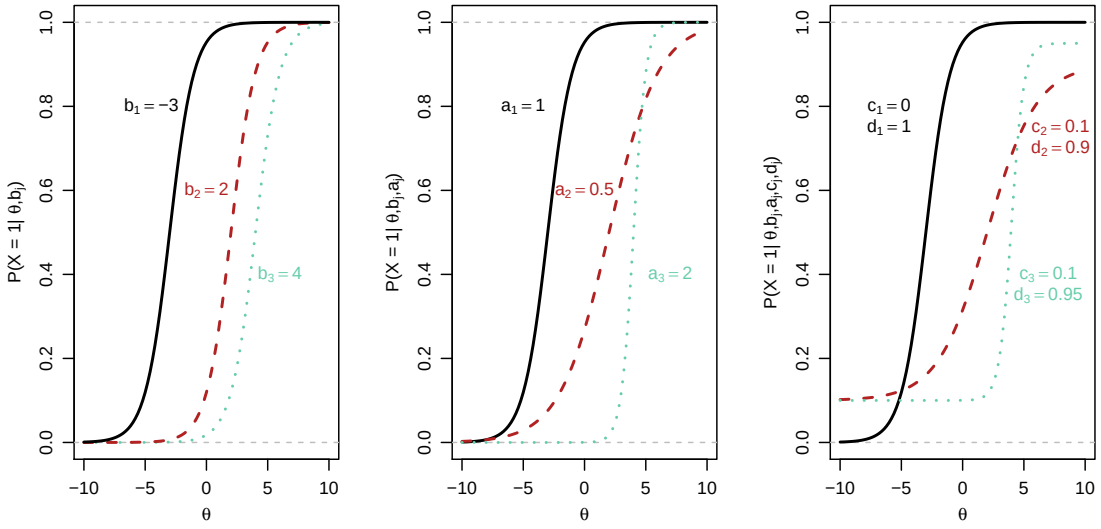

Figure 1. Item characteristic curves for three items according to the Rasch model (left), 2PL model (middle) and 4PL model (right) with varying parameters.

Extensions of the Rasch model include the two parameter model (2PL model, Birnbaum (1968)), which adds an individualistic slope for the ICC of each item a_j (Formula 2, Figure 1 (middle)). This means that some items can distinguish person abilities at different levels. The parameter a_j therefore represents an item's discrimination. The name of the model refers to the number of parameters relating to the items. Subsequently, the Rasch model can also be interpreted as a special case of the 2PL model with the constraint that $a_j = 1 \; \forall \; j = 1, \ldots, m$. Barton and Lord (1981) introduced another expansion, the four parameter model (4PL model), which expands the 2PL model further by adding two new item parameters. A lower asymptote c_j and a higher asymptote d_j in the ICCs (Formula 3, Figure 1 (right)). The inclusion of c_j with $c_j > 0$ makes it more likely that a person with a low θ score will answer item j correctly. It can therefore be interpreted as a guessing parameter. On the other hand, $d_j < 1$ lowers the probability of a correct answer

for highly skilled persons. It therefore represents the possibility of a slipping error. The 4PL model can be restrained as well with $c_j = 0$ *and* $d_j = 1 \; \forall \; j = 1, \ldots, m$, resulting in the 2PL model again.

$$P(X_{ij} = 1 \mid \theta_i, b_j, a_j) = \frac{e^{a_j(\theta_i + b_j)}}{1 + e^{a_j(\theta_i + b_j)}} \tag{2}$$

$$P(X_{ij} = 1 \mid \theta_i, b_j, a_j, c_j, d_j) = c_j + (d_j - c_j)\frac{e^{a_j(\theta_i + b_j)}}{1 + e^{a_j(\theta_i + b_j)}} \tag{3}$$

Formula 2 & 3: Probability of item j being answered correctly by person i under the 2PL (above) and 4PL (below) model respectively with θ_i representing the person's ability, b_j representing the item's easiness, a_j representing the item's discrimination, c_j representing item-specific guessing and d_j representing item-specific slipping errors.

Rasch (1966) acknowledged some of these expansions when introducing the Rasch model, but pointed out that the restrictions he chose, create some desirable attributes. The most notable being the sufficient statistic for θ_i which is delivered in the sum of correctly answered items given the item's easiness are known. The same is true for the b_js, where the number of people in a given sample that answer the items correctly is a sufficient statistic. From this, the conditional maximum likelihood (ML) estimation was derived (Andersen, 1970). It is possible within the constraints of the Rasch model and provides parameter estimation without further assumption as for example the marginal ML estimation has to make. The marginal ML estimation is a helpful tool for parameter estimations in higher level IRT models (models with more than one item parameter), because it bypasses the need for a sufficient statistic for θ_i. In order to do that, it assumes the distribution of the θ_i parameter (Bock & Aitkin, 1981). As the Rasch model can use conditional ML estimation, there is no need for this assumption.

For the same reasons, person parameters can be compared independently to the items and in turn the item easiness can be compared independently to the sample. Furthermore, the Rasch model requires smaller sample sizes as the specifications on item parameters obviates the need to estimate their values. A stable estimate of the fewer parameters in the Rasch model is possible with a small sample size. These desirable attributes form the key advantages of the Rasch model. On the other hand, the restrictions of the Rasch model amplifies the problem of model fit. More

parameter specifications lead to a less flexible model. The Rasch model is therefore more likely to experience poor model fit. Although this can also be caused by a variety of reasons, for example differential item functioning.

Differential item functioning

DIF describes different subgroups of a population having different probabilities of a correct answer for a specific item. It is to be distinguished from impact which refers to subgroups having a stable consistent difference in ability. Impact also leads to different probabilities of a correct answer between the groups. DIF occurs when the subgroups are matched with respect to the ability yet the probabilities still differ (Dorans, 1989). Impact is represented in the θ_i dependent on the group affiliation. It is therefore compatible with item response modelling. DIF represents an additional influence on $P(X = 1|\theta_i, b_j)$(in the case of the Rasch model) other that the given parameters θ_i and b_j, which violates the model's assumptions. Impact will not be further explored in the current study to keep the focus on DIF detection.

In the field, the true abilities of subgroups are not observable. This makes the distinction between DIF and impact a challenge for which many solutions have been proposed for various models and kinds of DIFs over the years (Andersen, 1973; Glas, 1999; Suárez-Falcón & Glas, 2003). Andersen (1973) proposed the Likelihood quotation test building on the Rasch model's property, that item easiness is not independent of a particular sample. It uses this to construct a test which compares the product of the Likelihood of θ in the considered subgroups with the Likelihood in the whole sample. The method presupposes awareness of the subgroups where DIF is supposedly occurring. This is a disadvantage that most methods for detecting DIF share. It is inarguable that there are possible and even likely scenarios in which DIF does occur concerning an unexpected covariable. An example would be IQ tests and the covariable of pre-existing experience with IQ tests. It is well known that most items in IQ tests are trainable. Therefore, the participant's experience should create DIF in the item's easiness, but this is not accounted for normally. The typically examined covariables such as age, gender and education are far from the only factors that can create DIF. Furthermore, traditional DIF tests like the

Andersen Likelihood quotation test need to know the exact cut point which creates the subgroups in the covariable. Age for example is divided by decades in most test manuals with no examination or evidence that the item functioning differs in parallel to the decade system (for example Costa Jr and McCrae (2008)).

The traditional DIF detection tests also fail to recognise more complex or not standard types of DIF (Strobl et al., 2015). A short insight into existing categories of DIF helps to understand this difficulty.

Categorizing DIF

DIF was introduced above as relating to groups, i.e. a covariable defines distinct groups which have different probabilities of a correct answer for an item. These groups can be clearly defined by categorical covariables. One example being gender and gender bias of tests. Subgroups can also be defined by continuous covariables as is often the case with age. Although there is no clear understanding if DIF in this case is functioning categorically or continuously, most methods for detecting DIF assume a categorical divide created by the continuous covariables. This distinction also leads to a modelling difference. When DIF is viewed categorically, the item easiness can be adjusted in each group, making DIF a function of the items. Continuous DIF can be described by a multidimensional model, meaning that more than one person parameter is modelled (Ackerman, 1992). DIF is then seen as a secondary trait of the person which is disproportional distributed in regard to a covariable, rather than an attribute of the items. Although continuous covariables were used for simulating DIF in the current study, DIF was simulated as if the cut point created a dichotomy. This was in parallel to the Rasch-trees understanding of DIF.

Because DIF is a function of the items, it effects the estimation of the item easiness parameter in the Rasch model. DIF can effect every item parameter. This is reflected in the simulated conditions. As Rasch-trees are based on the Rasch model, only the estimation of the item easiness parameter can be detected. Other parameters have specified values in the Rasch model and are therefore not estimated.

Data simulation under higher level IRT models provides the possibility of cross-

ing DIF. The name refers to the crossing ICCs of the relevant subgroups (Figure 2). This means, that for certain range of θ one group has a higher probability of answering the observed item correctly but a lower probability in other ranges of θ. Therefore, the difference between the probability of a correct answer in two groups changes sign (Li & Stout, 1996). It stands in contrast to uniform DIF, where the ICCs of the subgroups are parallel to each other, but offset on the dimension θ. Crossing DIF is to be distinguished from non-uniform DIF, referring to DIF where the ICCs of the subgroups are not parallel, but don't necessarily cross. Crossing DIF presupposes an intersection point between the curves. A cut point is guaranteed if there is no DIF in the item easiness parameter as both curves display a probability of .5 at their item easiness, creating an intersection point. Crossing DIF is therefore present when DIF occurs in the item discrimination, but not the item easiness parameter. This is not the only condition under which crossing DIF occurs, but this is how it will be conceived in the present study. Crossing DIF is of special interest, because it is seldom detected by traditional DIF tests. There are tests especially created for registering crossing DIF (Li & Stout, 1996), but this prompts the same difficulty as the selection of covariables. When crossing DIF is not suspected, these methods will not be used, making the detection of crossing DIF unlikely.

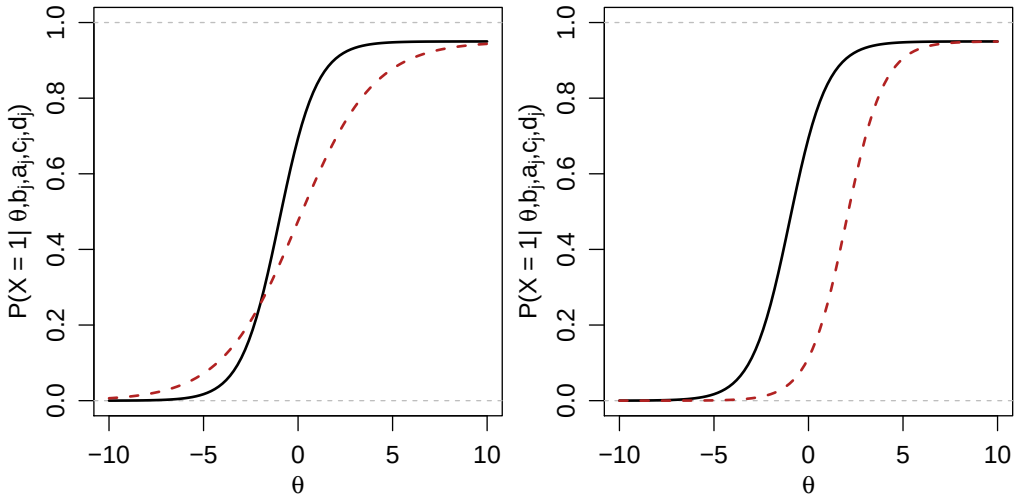

Figure 2. Item characteristic curves of one item and two groups showing crossing DIF (left) and uniform DIF (right).

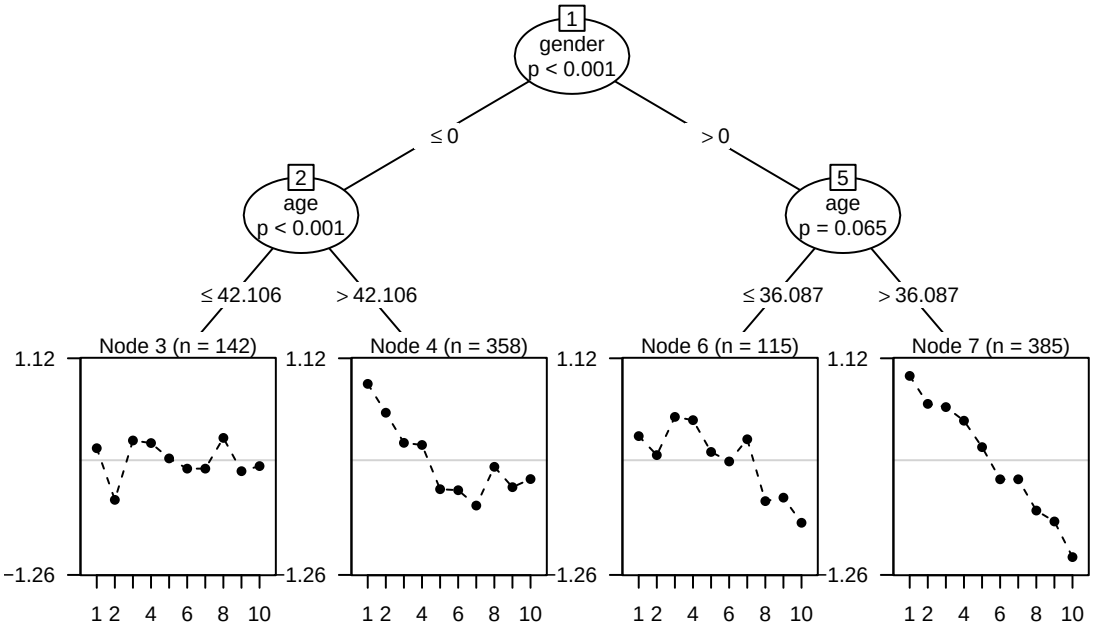

Figure 3. Example Rasch-tree with detected DIF in two covariables, resulting in four end nodes. The covariable age has different cut points depending on the gender subgroup.

Rasch-trees

A more flexible alternative to traditional DIF tests was introduced by Strobl et al. (2015) with the method of Rasch-trees. The goal was to be able to detect DIF without defining the cut points in covariables ahead of the analysis and also to provide results that are easy to interpret. This easy interpretation was implemented with a visualisation tool, which makes the tree structure visible (Figure 3). Rasch-trees are based on the principle of recursive partitioning of the data in all of the necessary subgroups. For that, the item easiness parameters are estimated in the entire sample. Then, the stability of the parameter estimation is considered for every available covariable. Finally, the sample is split in regard to the covariable which shows the most instability at the cut point that will improve the model fit the most. Two subgroups are created and the three steps are repeated in these subgroups. The process repeats until there is no significant difference found between further possible subgroups or until the number of observations in a resulting subgroup would be to small to ensure a stable estimation of the item parameters. How small this number of observations has to be is not set by the authors but left for the user of Rasch-trees to decide.

For the instability analysis in step two, the individual deviation from the parameter estimation with the entire sample is ordered regarding each of the available coavariates separately (Formula 6). If DIF is present in respect to a covariable, the ordered deviations should show a systematic change in respect of this covariable. Strobl et al. (2015) then use two different test statistics depending on the relevant covariable. Formula 5 is used for numerical variables and is limited by the suprenum of a tied-down Bessel process. For categorical variables, Formula 4 is used and limited by a χ^2-distribution. On each level, every possible covariable is tested in this way and the next sample split occurs at the covariable with the smallest p-value.

$$S_l = max_{i=\underline{i},\ldots,\bar{i}}(\frac{i}{n} \cdot \frac{n-i}{n}^{-1})\|W_l(\frac{i}{n})\|_2^2 \tag{4}$$

$$S_l = \sum_{q=1}^{Q} n(\sum_{i=1}^{n} I(x_{il} = q))^{-1}\|\Delta_q W_l(\frac{i}{n})\|_2^2 \tag{5}$$

$$\text{with} \quad W_l(t) = \hat{\mathbf{V}}^{-\frac{1}{2}} n^{-\frac{1}{2}} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \Psi(\mathbf{u}_{(i|l)}, \hat{\boldsymbol{b}}) \quad (0 \leq t \leq 1) \tag{6}$$

Formula 4, 5 & 6: Test statistics measuring the estimation instability with $i = 1, \ldots,$ n numbering the observations consequently to the lth covariable, $q = 1, \ldots,$ Q marking the qth hypothetical category, t defining a fraction of the sample, Δ_q the increment within the qth category and $\hat{\mathbf{V}}$ representing the outer-product-of-gradients estimate of the covariance (Strobl et al., 2015).

After selecting the covariable, the cut point is determined. Item parameters are estimated for the hypothetical of every possible cut point. The conditional ML estimation is used for this. The estimations for both groups are added within and between the groups (Formula 7) and the cut point which maximises this sum is chosen. This creates two subgroups in which the process is repeated until one of the stop criterion is reached.

Strobl et al. (2015) present many desirable attributes of Rasch-trees in their

$$\sum_{i \in L(\xi)} \boldsymbol{\Psi}(\mathbf{u_i}, \hat{\mathbf{b}}^{(\mathbf{L})}) + \sum_{i \in L(\xi)} \boldsymbol{\Psi}(\mathbf{u_i}, \hat{\mathbf{b}}^{(\mathbf{R})}) \tag{7}$$

Formula 7: Sums of the conditional ML estimation over every person i in the two subgroups created by a hypothetical cut.

introductory paper. In regard to the Type-I-Error, the trees are fairly conservative, the Power only reduces meaningfully when the cut points reach the periphery in a covariable and it detects more complicated DIF forms more reliably than traditional tests such as the Andersen likelihood test. Furthermore, the results are easy to interpret and the cut points are detected instead of given in advance, which are the main goals of Rasch-trees. Besides this very detail oriented approach, Rasch-trees can also be used as a test of overall item fit. General person-free estimation of the item easiness implies the same estimations for every possible subgroup with every possible cut point. The Rasch model therefore applies when no split is performed with the Rasch-tree method resulting in only one node. This does not mean that the Rasch model applies in every subgroup created by Rasch-trees as the stopping criteria for splitting also includes factors like a sample size, which is too small.

The simulation studies in Strobl et al. (2015) of course assume the intended use case for Rasch-trees, i.e. DIF detection under the specifications, that the Rasch model generally applies. The purpose of the current work is to investigate the behaviour of Rasch-trees when the Rasch model no longer applies. More specifically, a variation of DIF conditions are simulated under the general framework of the 2PL or 4PL model. While the Rasch model is the most used and the most comprehensible IRT model, it is also the most restrained, making its specifications and implications varying degrees of unrealistic in the field. A multiple choice test for example has a clearly defined guessing probability for the correct answer regardless of ability. Especially with dichotomous items as used here, the probability for guessing the correct answer is high. The Rasch model however does not account for guessing. This influences parameter estimations, making them less accurate. In this context, the assumption that a person will answer an item accordingly only to her true ability is unrealistic as well as hindering to the estimation of people's ability (Liao, Ho, Yen, & Cheng, 2012). The 4PL model would be more applicable.

This recommendation prerequisites the wide spread acceptance and usage of IRT models, which is not the case. Most test manuals still use classical test theory and only mention IRT and the Rasch model briefly (for example Costa Jr and McCrae (2008)). Research in the differences and extensions of IRT models is without

any doubt important and should be fostered right now, yet comprehensive usage of appropriate IRT models is a long way to go still. In the mean time, it is relevant to discuss the effects of model misspecifications, i.e. the Rasch model's usage in cases where it does not apply. To this end, a series of simulations were performed, all constructing data either under the 2PL or the 4PL model with different levels of DIF. The constructed data were then examined with the Rasch-tree method and results were compared. The study was completely explorative with overarching thematic questions rather than specific hypothesis. The most prominent question being how much of a problem the fitting of the data to a higher level IRT model causes for the DIF detection. This is an extension of the question, how important the specification of the "right" IRT model is in general, which will be considered more thoroughly in the discussion.

Can DIF detection still reliably succeed even with model misspecification or is the selection of the right IRT model a necessary prerequisite for any DIF investigation? In the field of course, no data set perfectly corresponds to one model, IRT or otherwise. The data in this study is created from a specific model while observed data can only be described adequately by any model. Nevertheless, simulations can give a more unambiguous insight than field studies and offer an estimation on how catastrophic wrong assumptions and specifications can be in this particular matter. Due to the explorative nature of the study, there were no concrete hypothesis. Instead, a focus was set in observing patterns in the gathered data with no confirmatory statistical analysis.

It was not expected that the Rasch-trees show comparable results as in Strobl et al. (2015). A decrease in reliability and general performance is to be expected as the Rasch-tree method was not build for the circumstances presented in the present work. The current results are therefore not a critic in any kind, but an examination of possible scenarios.

Methods

All simulations were implemented in the R system for statistical computing (R Core Team, 2019). The Rasch-tree method was conducted with the package "psychotree" (Strobl et al., 2015).

Simulation procedure

In order to simulate data from different models with or without DIF and gather the criterion variables, the following steps were performed.

1. A standard normally distributed vector of person parameters was specified. Neither parameter space nor ability distribution are constrained in IRT models, but a normal distribution is often assumed.

2. A matrix P containing the probabilities of a correct answer was computed with every cell p_{ij} defined by equation 2 or 3, depending on the used model. The necessary item parameters were either set overall (b_j) or by the simulated condition (a_j, c_j, d_j). If DIF was simulated, P was manipulated according to the specifications.

3. The responses were drawn from a binomial distribution with n = 1 using the matrix P as probabilities.

4. The Rasch-tree method was applied including the estimation of the item easiness parameter in the end nodes under the Rasch model. The α-level for the decision of DIF detection was set to .1. Rasch-trees tend to be conservative in this decision as Strobl et al. (2015) demonstrated.

5. The number of end nodes, significant results and the RMSE were collected.

Experimental settings

A total of 516 conditions pooled into 17 groups were simulated with varying factors while the following terms remained the same over all conditions.

- Number of replications.

  For every condition, 500 replications were performed to balance accurate estimations of the reported values and the limited computing power available.

- Number of observations.

  Each replication created a sample size of 1000 observations.

- Number of items.

  A total number of 10 items were simulated in each condition. The simulated easiness parameter were chosen as $b = \{-2, -1, 56, -1.11, -0.67, -0.22, 0.22, 0.67, 1.11, 1.56, 2\}$ in order evenly cover the whole ability parameter spectrum and still ensure stable simulations for all easiness parameters. A Person-item map showcasing $\hat{b}$ and $\hat{\theta}$ of one sample on the same latent dimension is shown in Figure 4. If the item discrimination was not explicitly differed, it was set to $a = \{.60, .64, .69, .73, .78, .82, .87, .91, .96, 1\}$ to reflect a cohesive test with varied item discrimination. If the guessing parameter was not explicitly differed, it was set to $c = \{0, .03, .07, .10, .13, .17, .20, .23, .27, .30\}$. The same was true to the parameter representing the slipping error, which was set to $d = \{1, .97, .93, .90, .87, .83, .80, .77, .73, .70\}$ if not explicitly differed.

The following terms where varied over the simulated conditions.

- Number of item parameters.

  The baseline in the condition with no simulated DIF is the Rasch model with only one item parameter representing easiness. All conditions that were simulated under the 2PL model accordingly had two item parameters, easiness and discrimination. The 4PL model adds two more representing guessing and slipping errors. Therefore, data simulated under the 4PL model has four item parameters.

- Magnitude of DIF.

  The magnitude of DIF was varied in accordance to the parameter in which DIF was simulated. For a_j the range from 0 to 1 was chosen. For c_j and d_j

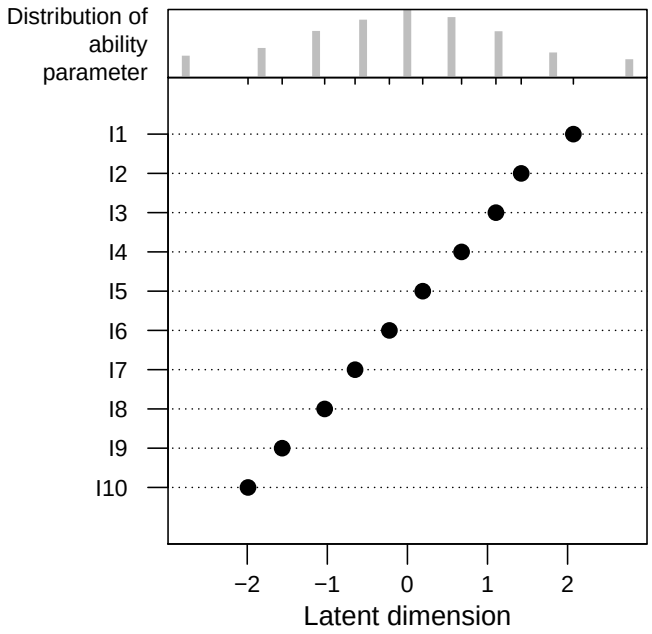

Figure 4. Person-item map with the distribution of the ability parameter in one sample and the ten item easiness parameters depicted on one latent dimension.

the range from 0 to .5 was chosen. Bigger DIF than .5 in these parameters would be unrealistic and could lead to a large percentage of observations with only zeros or ones.

- Number of items with DIF.

  Every variable constellation was simulated concerning 0, 2, 5 or 8 number of items. Conditions with zero influenced items are used as a control inside every condition. The other three levels are representative for the inclusion of DIF in few, half and many items of an entire test.

- Parameter(s) effected by DIF.

  DIF was always simulated in the parameter(s) new to the model, meaning the discrimination for the 2PL model and the guessing and slipping error parameter for the 4PL model together. There were conditions in which DIF was simulated in the item easiness parameter as well as conditions were DIF only affected the item easiness parameter for the comparison of uniform and crossing.

- Covariable determining DIF.

  As was stated in the introduction, the Rasch-tree method uses different test statistics for detecting groups in categorical or continuous covariables (equations 4, 5). DIF was therefore either simulated in categorical and continuous covariables. The categorical covariable was dichotomous. The continuous was drawn from a uniform distribution with a range from 18 to 100 (in analogue to age as an often used covariable). Here, the cut point for the subgroups had to be determined, leaving more room for estimation errors. The simulated cut point was chosen towards the middle of the variable at 40 in order to avoid loss of Power which can occur with cut points towards the edges of a covariable (Strobl et al., 2015). Although the covariable was continuous, DIF was simulated as if the cut point created a dichotomy. Finally, there were conditions as well in which DIF was simulated in regard to both covariables.

Criterion variables

To evaluate the impact of the simulated conditions, the following variables were conducted and compared. No standard values were set for the criterion variables. Every condition contained a simulation where no items contain DIF which was used as a baseline to compare the other results to.

- Significant results.

  The number of significant results indicates how often DIF was found. This either represents the Power of the Rasch-trees as a global DIF test, when DIF was simulated, or the Type-I-Error rate, when it was not.

- Number of end nodes.

  A feature of Rasch-trees is, that they not only indicate that DIF was found, but also in how many and which covariables as well as the number and placement of cut points. The number of end nodes, which represent the subgroups in which no further split was carried out, therefore provides further information on the detected DIF. In case of conditions where no DIF was simulated, this metric not only shows the number of false positives, but also how far off the

method was on average. In this case, the heavy influence of extreme data points is welcomed, as these could lead to highly wrong conclusions, even if rarely.

- RMSE.

  The root mean square error of the item easiness parameter was reported. It gives a representation of the estimations accuracy for the item easiness parameter.

Results

Simulations without DIF

Figure 5 shows the influence of rising a_j, c_j and d_j parameters respectively on the probability of a significant result (top), the number of end nodes in the Rasch-trees (middle) and the RMSE (bottom) without simulated DIF. In the left column (graphs 1), the item discrimination a_j was varied from 0.1 to 1 in either zero, two, five or eight items. The other items had a fixed item discrimination of 1. In the middle column (graphs 2) , the guessing parameter c_j was varied from 0 to 0.5 in either zero, two, five or eight items. The other items had a fixed guessing parameter of 0 and all items had a fixed slipping error parameter d_j of 1. The slipping error parameter d_j was varied inversely to the guessing parameter, depicted in the right column (graphs 3).

Neither the Type-I-Error nor the number of end nodes are affected by any of the new parameters, impartial of the number of items diverging from the Rasch model restriction ($a_j = 1$, $c_j = 0$, $d_j = 1 \; \forall$ j = 1, ..., m). The RMSE is influenced by both independent variables for all three new item parameters. With rising a_j values, the RMSE decreases, indicating a more accurate estimation of the item easiness. Notable are the conditions with a_j values .9 or 1. Here, the RMSE falls below the conditions where no items diverge from the restriction $a_j = 0 \; \forall$ j = 1, ..., m. A rising c_j-value leads to a higher RMSE, the same goes for a sinking d_j-value. The change in these two parameters lead to a smaller increase in comparison to the influence of

the item discrimination. The influence of all three new item parameters are bigger, the more items had new parameters diverge from the Rasch model restrictions.

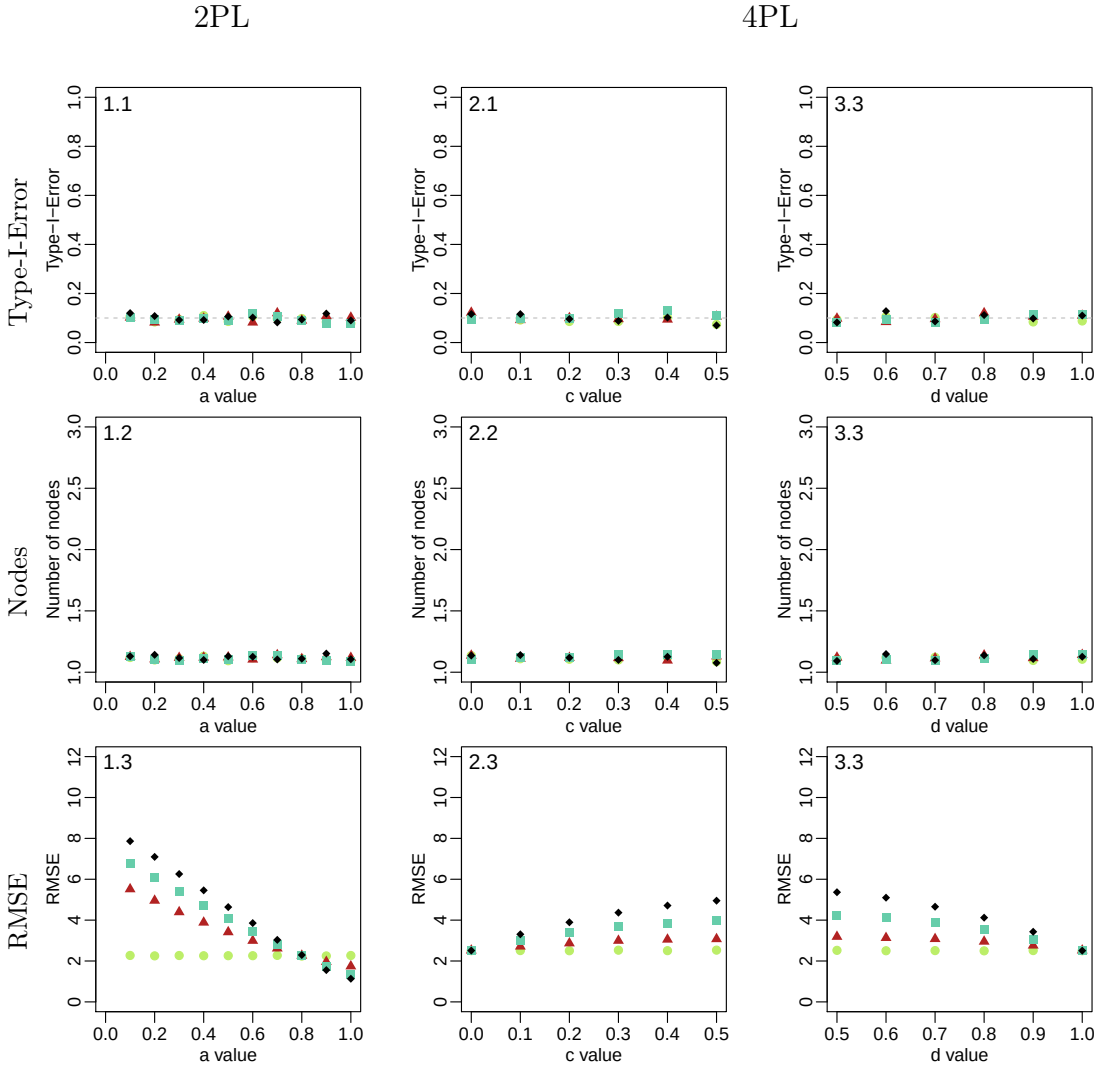

Figure 5. Type-I-Error rate (top), Number of end nodes in the Rasch-trees (middle) and Root mean square error (RMSE, bottom) in the end nodes dependent on the values of a_j (left), c_j (middle), d_j (right) respectively without simulated DIF in these parameters (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items with divergent parameter values from $a_j = 1$ or $c_j = 0$, $d_j = 1$ $\forall$ j = 1, ..., m).

The main results (graphs 4-11) are presented separately for the three criterion variables in order to facilitate comparisons and create a comprehensive overview.

Probability of a significant result

Figure 6 illustrates the influence of the different simulated conditions on the probability of a significant result in the Rasch-trees. This shows the Power of the

Rasch-trees as a global DIF test when it was simulated and the Type-I-Error rate when no DIF was simulated. The top four graphs depict simulations under the 2PL model, while the bottom four graphs show simulations under the 4PL model. In each case, the top two graphs have no simulated DIF in the item easiness parameter b_j, while the conditions depicted in the bottom two graphs include a simulated DIF according to the same covariable that influences the new parameters with a fixed value of 0.5 in b_j. This means, the focal group has a higher item easiness of 0.5 in comparison to the reference group. Generally, the left side graphs include simulated DIF in regard to the categorical, dichotomous covariable. The right side graphs have simulated DIF in regard to the continuous covariable, ranging from 18 to 100 and cut at 40. A horizontal line was added at .1, marking the α-level.

All conditions where no item is affected by DIF (green dots) serve as a baseline to compare the other conditions to. They are located at the set α-level over all conditions. The Type-I-Error rate does not inflate regardless of the used model. The same goes for the conditions in which no DIF in the item easiness is simulated (graphs 4.1, 5.1, 8.1, 9.1) and the DIF value for the new parameters is 0. There is no DIF to detect in these conditions. Therefore, the probability of a significant result unitarily drops to the α-level. Notably, when DIF is simulated in the item easiness, but not in the new parameters of the 4PL model (graphs 10.1, 11.1), the probability of a significant result stays at the α-level as well.

When the DIF value in the new parameters is rising, the probability of a significant result does not rise quicker in graph 10.1 than in graph 8.1, where no further DIF in the item easiness is present. This is the case for the conditions regarding the categorical covariable. In the conditions with the continuous covariable under the 4PL model, the probability of a significant result rises quicker in graph 11.1 than in graph 9.1. It is also notable, that in all graphs of the 4PL model, the points of the conditions with two, five and eight items affected lie on top of each other under otherwise identical conditions.

The Rasch-trees detect DIF in the b_j under the 2PL model (graphs 6.1 and 7.1). Here, when the DIF value for a_j is 0, the probability of a significant result is higher than the α-level. As there is no DIF in a_j, the Rasch-trees seem to detect the

DIF in b_j. However, the probability is unanimously low, almost at the α-level, for the DIF value of 0.1. In general with the 2PL model, the probability of a significant result rises almost logarithmically with rising DIF value.

Number of end nodes

Figure 7 is structured in the same way as Figure 8 and shows the influence of the DIF value and number of affected items in different conditions on the number of end nodes in the Rasch-trees. When no item is affected and there is no DIF in b_j, the average number of end nodes in the Rasch-trees should be slightly above 1. When DIF is simulated in any way, there should be slightly below two nodes, as both covariables create two groups. Here, both DIF for b_j and for the new item parameters in each model are simulated in parallel to the same groups. The number of end nodes and the probability of a significant result are directly dependent on another, but the number of end nodes provides the additional information of how many splits are made when DIF was detected.

Because of this dependency, some of the patterns are parallel to the results in figure 6. All conditions with zero items affected by DIF show a little over 1 node on average. The points under the 4PL model and with more than zero items affected also lay on top of each other. In the 2PL model, the conditions with two and five items affected performed similarly, while the conditions with eight affected items on average have more nodes at a smaller DIF value. DIF in the new parameters of the 4PL model (graphs 8.2-11.2) generally have less of an impact on the number of end nodes than DIF in the item discrimination under the 2PL model (graphs 4.2-7.2). The DIF in b_j under the 4PL model (graphs 10.2 and 11.2) is not detected as often as under the 2PL model (graphs 6.2 and 7.2).

More notable is the upper asymptote which all conditions under the 2PL model (graphs 4.2-7.2) and the conditions in graph 11.2 under the 4PL approach. No condition has more than 2.2 nodes on average even if the Power approaches 1 in the same conditions (4.1-7.1 and 11.1). A line was added at 2.2 to highlight the asymptote. It is reached at the same DIF values as the Power asymptote of 1 in figure 6. Even high values of DIF in the new parameters of the 4PL model does not

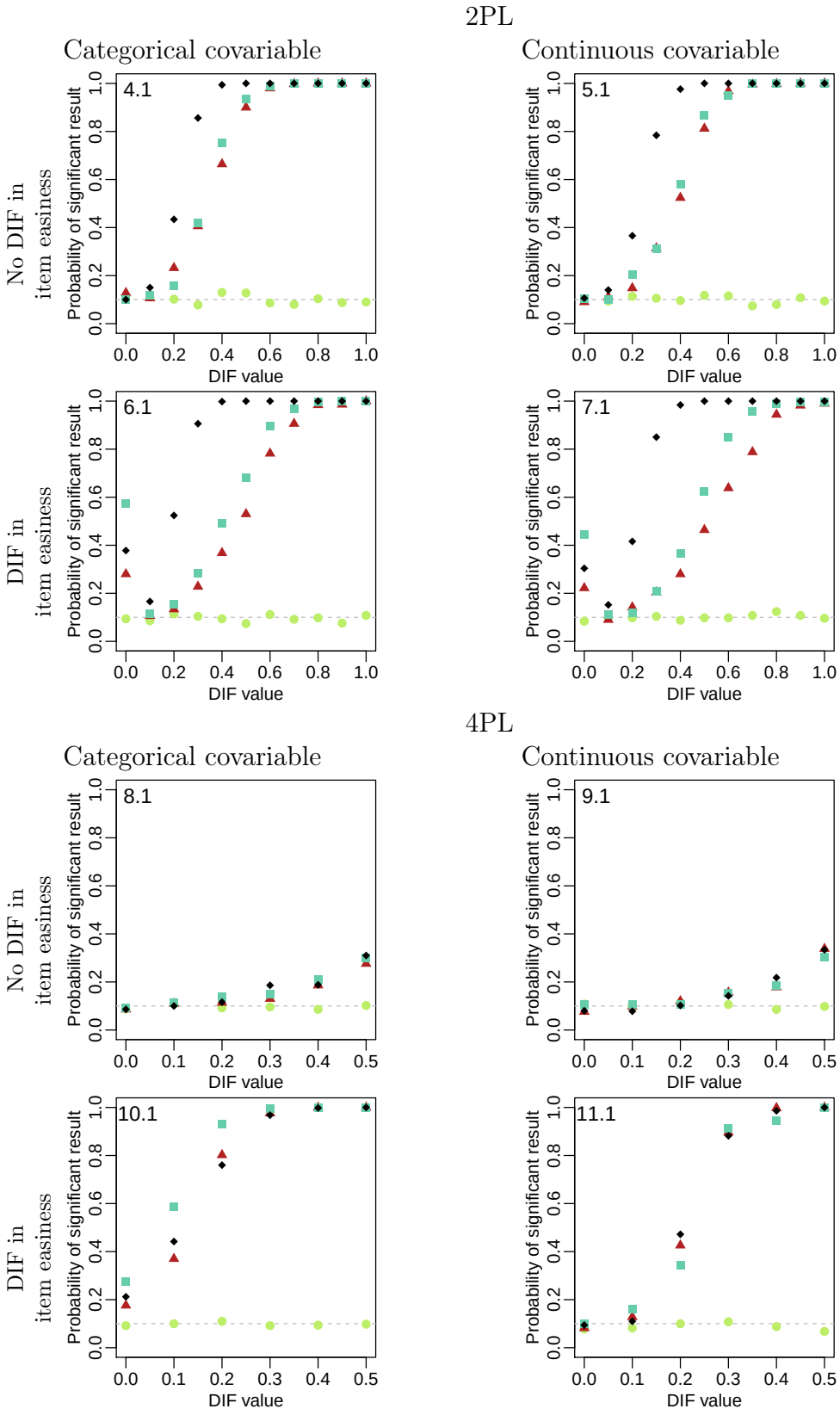

Figure 6. Probability of a significant result in the Rasch-trees dependent on the scale of DIF and the number of items affected by DIF (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items affected).

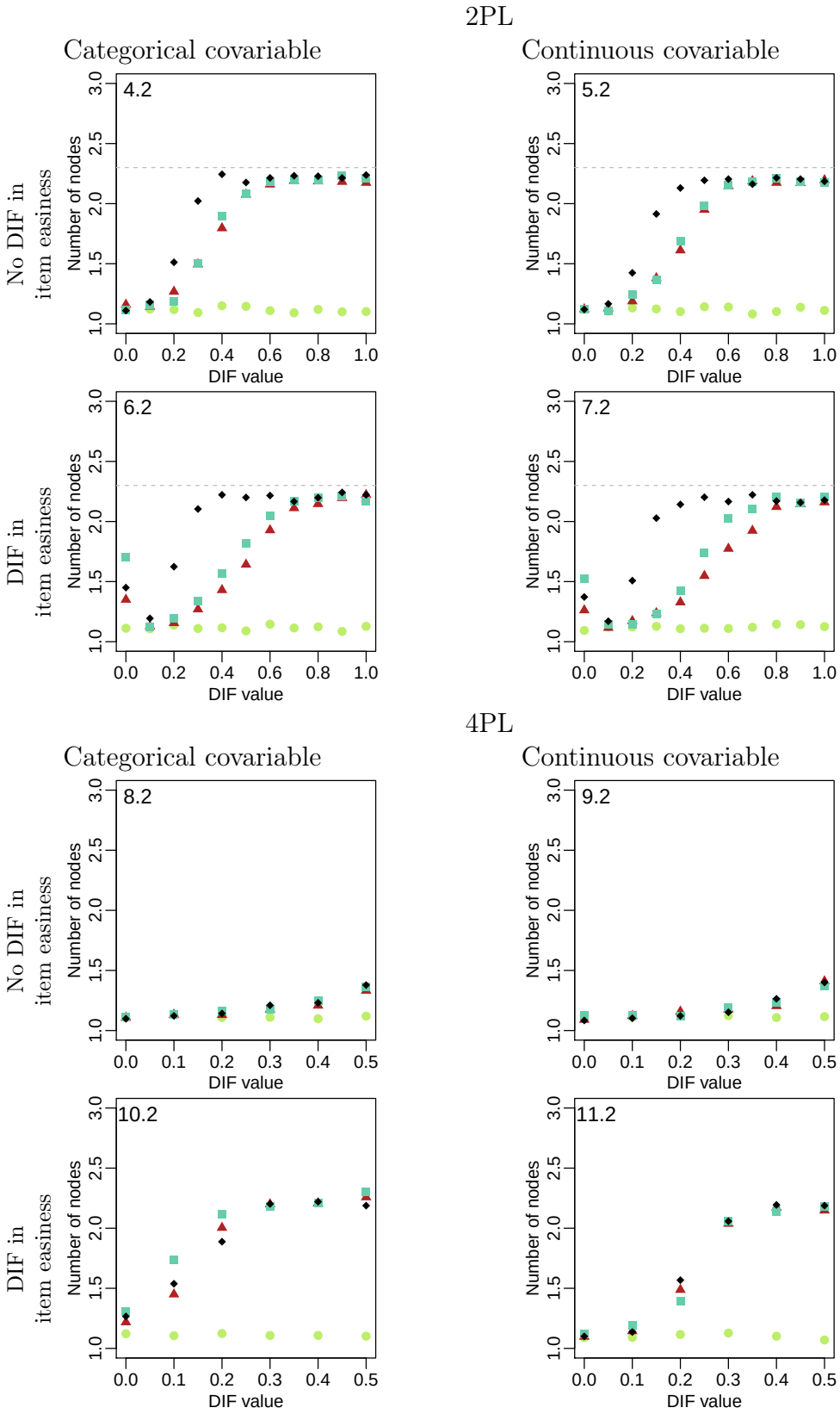

Figure 7. Number of end nodes in the Rasch-trees dependent on the scale of DIF and the number of items affected by DIF (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items affected).

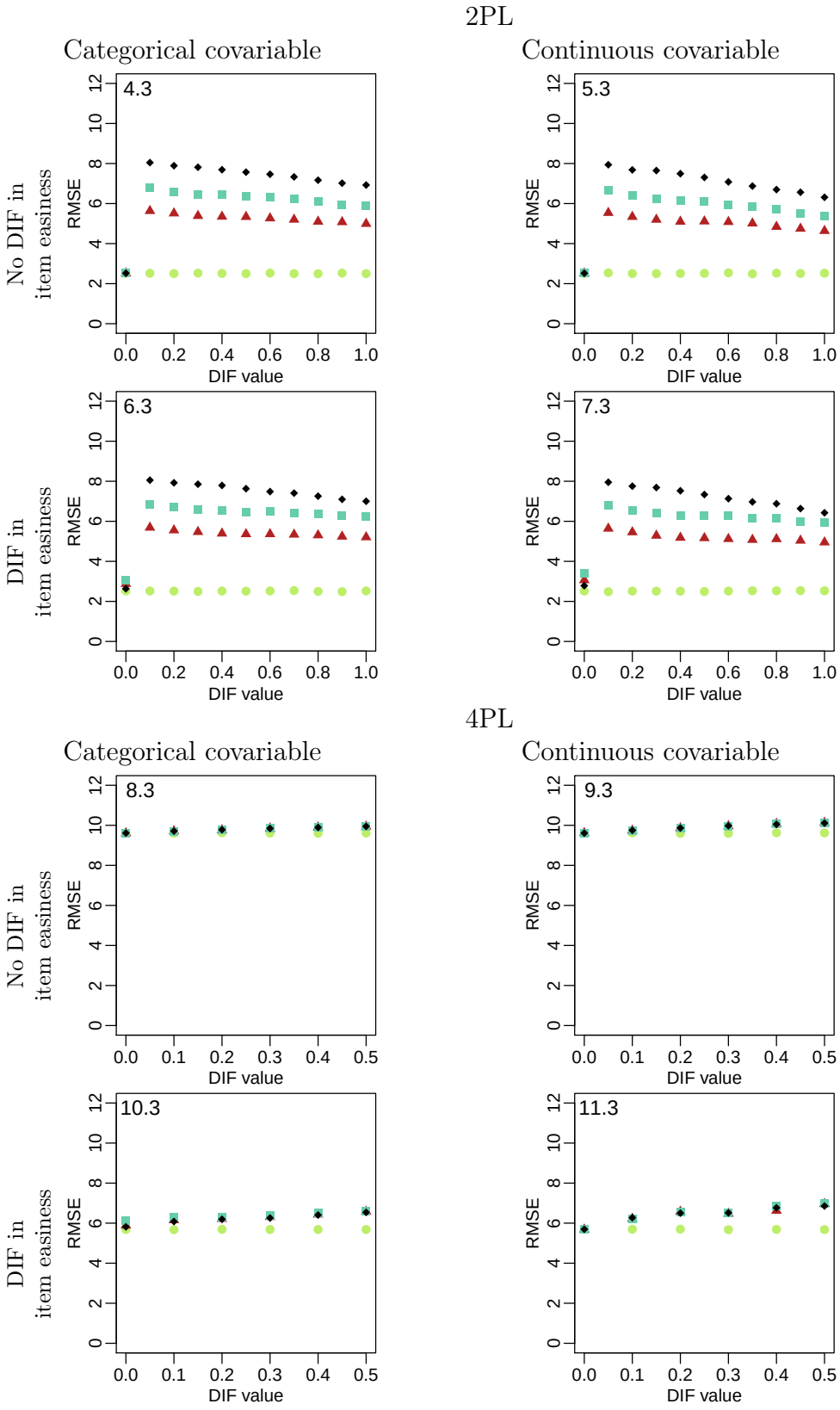

Figure 8. Root mean square error (RMSE) of the item easiness estimations in the Rasch-tree end nodes dependent on the scale of DIF and the number of items affected by DIF (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items affected).

raise the number of end nodes largely except for graph 11.2.

Root Mean Square Error

Figure 8 is structured in the same way as Figure 6 or Figure 7 and shows the influence of the DIF value and number of affected items in different conditions on the RMSE in the end nodes.

The conditions with a DIF value of 0 show a similar pattern as could be observed in the probability of a correct answer and the number of end nodes. They lay on top of each other except in the graphs 6.3, 7.3 and 11.3. The RMSE is not especially affected by the DIF value or the number of affected items under the 4PL model. Only in graph 11.3, when DIF in the continuous covariable is affecting b_j, c_j and d_j, correspond higher DIF values to slightly higher RMSE. This is not dependent on the number of affected items. Notably, the conditions with no affected items have an increased RMSE under the 4PL model as well. Even higher RMSE can be observed in the graphs 8.3, 9.3, 10.3. Here, neither the DIF value nor the number of affected items influences the RMSE. It is unitarily located at an increased level.

For the 2PL model, the conditions with no affected items were on the baseline level seen in the simulations without DIF. The more items were affected, the higher was the RMSE, but the higher the DIF value, the lower the RMSE. This relation between DIF value and RMSE was more distinct in the conditions with the categorical covariable than in the conditions with the continuous covariable. Whether DIF was present in b_j did not effect this pattern.

Crossing DIF

The conditions where DIF was simulated in the item discrimination, but not in the item easiness have intersecting ICCs of the resulting subgroups, thus creating crossing DIF. These conditions were already observed in the graphs 4, 5, 8 and 9. Figure 10 and 11 in the Appendix compare them with uniform DIF, where only the item easiness is affected by DIF. Figure 9 in the Appendix shows the ICCs of one data set out of the simulations with uniform and one with crossing DIF in the

categorical covariable. The Rasch model still does not apply here as a_j is not set to 1 in all items. There are no notable differences in any criterion variable between the conditions with uniform or crossing DIF.

DIF in two covariables

In all previous conditions, DIF was simulated in regard to the categorical or the continuous covariable, but only one at a time. The following conditions (Figures 12 and 13 in the Appendix) were simulated with DIF in regard to both the categorical and the continuous covariable. This means that more than two subgroups were created. The DIF value of the continuous covariable was kept constant, while the DIF value of the categorical covariable was varied as seen before. Also like before, DIF was always affecting a_j under the 2PL model and c_j, d_j under the 4PL model. In half of the conditions, b_j was affected as well in regard to the continuous covariable with a constant value of 0.5. Because the DIF value in the categorical was varied, the results are best compared to the graphs 4,6,8 and 10, where DIF in the categorical covariable was varied, but no DIF in the continuous covariable was present. Thereby, graph 16, where DIF was affecting a_j and b_j, is unremarkable compared to graph 6. The general rise of Power is similar in 14.2 compared to 4.2, although it begins at a lower DIF value. Although more subgroups were simulated, the number of end nodes does not rise. To the contrary, a higher DIF value is needed and the upper asymptote is lowered in 14.1. In contrast to earlier results, the conditions under the 4PL level show a rise in Power which starts at a lower DIF value than under the 2PL model. These conditions now also reach the upper asymptote of 2.2 in regard to the number of end nodes. The RMSE is unaffected by the simulation regarding two covariables.

Discussion

For the current work, various simulations were performed in order to assess the effects of model misspecification, more specific the false assumption that the Rasch model applies to data for detecting DIF with the Rasch-tree method. The

overarching research question was whether DIF detection can still be achieved even if the restrictions of the Rasch model are not met. The results show, that the specifics of DIF, model parameters and relevant covariable are needed to answer this question.

DIF in the item easiness parameter, which should be detected reliably with the Rasch-tree method, is more likely to be detected when the data follows the 2PL model in comparison to the 4PL model. The 4PL model also results in a less accurate estimation of the item easiness. Without DIF, the RMSE was generally higher, the more and the further the item parameters stray away from the Rasch model. As the Rasch-trees assume the Rasch model for their estimations, this was to be expected. The Type-I-Error rate is not inflated regardless of the used model and even when DIF was present in more than one covariate. When DIF was simulated in the item easiness, but not in the new parameters of the 4PL model (graphs 10.1, 11.1), the probability of a significant result is also at the α-level. It should be higher as DIF is present in the item easiness, which the Rasch-trees should be able to detect.

While additional DIF in the item easiness parameter leads to a faster rise of Power under the 4PL model (graph 11.1) compared to the same conditions without DIF in item easiness (graph 9.1), the same pattern can not be observed under the 2PL model. Comparing graph 7.1 and graph 5.1, the Power with eight affected items is roughly the same, but rises at a later DIF value in graph 7.1 when two or five items are affected. The addition of DIF in the item discrimination has opposing effects under the 2PL compared to the 4PL model. This can also be observed with categorical covariables in the graphs 4.1 and 6.1 compared to 8.1 and 10.1, although less pronounced. DIF in the item discrimination parameter seems to be detected as DIF in the item easiness parameter so that the addition of DIF in the item easiness does not further the detection any more. This presumption could also explain the lack of difference in uniform and crossing DIF and could be a unique property of Rasch-trees. Most DIF tests either underperform with crossing DIF or were specifically constructed for it. For a test that was not designed for it, the lack of Power loss is certainly an asset.

The confounding of DIF in item discrimination for DIF in item easiness is

further supported by the upper asymptote observed for the number of end nodes, which seems to be independent of DIF in item easiness. The assessment, that there are not more than two groups even when DIF was present in more than one item parameter, was correct and perhaps trivial for the categorical covariable. The continuous covariable however yields the potential for systematically detecting more than two groups. This is not the case. The upper asymptote is located around 2.2 for all graphs under the 2PL model and graphs under the 4PL model where DIF was simulated in regard to two covariables. The value of 2.2 as opposed to 2, which should be the theoretical asymptote, can probably attributed to the Type-I-Error. It seems that no condition produces an exceptional number of outliers, which would drag the asymptote upwards. Therefore, no condition has drastically higher consequences in the form of identifying extremely more subgroups sometimes. On the contrary, even when more than two groups should be detected, the upper asymptote does not move upwards.

DIF in the new parameters of the 4PL model seems not to be mistaken for DIF in the item easiness so that the addition of DIF in the item easiness accelerates the rise of the Power, at least in regard to the continuous covariable (graphs 9.1, 11.1). It is further accelerated when DIF is present in regard to both covariables (graphs 15.1, 17.1). This can also be observed in regard to the number of end nodes and the RMSE. For the number of end nodes, even high values of DIF in the new parameters of the 4PL model does not raise the number of nodes largely except for graph 11. The RMSE as well seemed not especially affected by the DIF value or the number of affected items under the 4PL model. This stands in slight contrast to the effect changing c_j and d_j parameters had in the simulations without DIF. However, the RMSE is in general higher under the 4PL model, which is in agreement with the simulations without DIF.

The most straight forward explanation for this is that the Rasch-tree method actually recognises the difference between c_j, d_j and b_j better than the difference between a_j and b_j. This is supported by the behaviour of the Rasch-trees in regard to the continuous covariable, but contradicted by the behaviour in regard to the categorical covariable. A distinction between the parameters and DIF in them seems

only present in regard to the continuous covariable. Another possibility is that the parameter estimation under the 4PL model is not accurate enough to detect DIF. This is supported by the raised RMSE values under all 4PL conditions, but contradicted by graph 11.1, where the DIF detection seems to be only slightly less sensitive for DIF in b_j than the 2PL equivalent in graph 7.1. Then again, neither of these reasons can explain that the number of nodes and the Power rise when DIF was simulated in regards of two different covariables. This rather suggests that the simulated DIF was not severe enough in c_j and d_j to be detected.

Another curiosity in regard to the probability of a correct answer are the graphs 6.1 and 7.1, where the conditions with a DIF value of zero have a probability of a significant result above α. This was to be expected, as DIF is still present in the item easiness parameter. Curious are the conditions with a DIF value of 0.1. Here, the probability unitarily drops to the α-level. A similar abnormality can be observed in graph 1.3. The RMSE with a_j values of .9 or 1 are lower when two, five or eight is lower compared to the conditions when no item is affected. Especially with an a_j value of 1, the RMSE should be equal because not affected items had a set a_j of 1 as well. When all items have the same discrimination level and nothing else is altered, it is unclear where this discrepancy originates. The RMSE also behaves contraintuitively when simulating DIF under the 2PL model. The more items were affected, the higher the RMSE. However, the higher the DIF value, the lower is the RMSE. This relation between DIF value and RMSE is more distinct in the conditions with the categorical covariable than in the conditions with the continuous covariable. Whether DIF is present in the item easiness does not effect this pattern, neither does the simulation of DIF in regard two both covariables. This is an anomaly that can not be explained with the current results.

Stability of the Rasch model

This discussion on the detection of DIF is an expansion of the general discussion whether the Rasch model provides stable estimations when the data corresponds more to a higher level IRT model. There is no real agreement on this issue. While Dinero and Haertel (1977) argue, that the Rasch model is resilient at least consider-

ing deviations from $a_j = 1 \; \forall \; 1, ..., m$, while Skaggs and Lissitz (1986) identify it as a notable risk factor. It interfered in the parameter estimation more severely than a deviation from the specification $c_j = 0$. A similar observation can be made for the current data, as the 4PL model increases the RMSE more notably and mostly independent from DIF.

Forsyth, Saisangjan, and Gilmer (1981) pointed out that IRT models in general have strong assumptions, which are not always met. They argued, the general IRT model assumptions should be considered before the additional assumptions of the specific models and stresses the examination of unidimensionality and local stochastic independence of the items. Their results suggest, that the Rasch model is more robust against violations of these two assumptions than expected. Particularly violating unidimensionality did not affect the model fit extremely. Then again, a significant stint in the model fit was found by Slinde and Linn (1979) when examining violations of unidimensionality, local stochastic independence and power test. The power test assumption is relevant, because it implies that every question has the same probability of being processed. If this was not the case, as in speed tests, the probability of a correct answer would be dependent on not only the person and item parameters, a core assumption of IRT models. None of these assumptions were actively manipulated in the current simulations, but rather they were all unitarily met. Consequently, they could not interfere with the current results, but should in the future be accounted for actively in the light of the earlier comment, that DIF is sometimes interpreted as a second dimension for the person parameter.

Slinde and Linn (1979) emphasise again that higher level IRT models need a greater sample size and test length to achieve acceptable fit. The drop caused by their violations was therefore not compensable by using higher level IRT models with the same sample size and test length. The applicability of the Rasch model with small sample size is an advantage also stressed by van de Vijver (1986). They came to the conclusion that the Rasch model is a viable alternative for small sample sizes and test lengths even if $b_j \neq 1$ and $c_j > 0$ apply. This claim was backed by them finding hardly any diminishing of estimation accuracy when item discrimination was raised. The simulations at hand, which did not include DIF but altered the

item discrimination stands in contrast to this observations. In van de Vijver (1986), higher guessing parameter only had an impact on the variance of the estimations not the accuracy as well. This is consistent with the current simulations with varying guessing parameters, but in slight contrast to the RMSE in the conditions with DIF, where the RMSE was consistently raised.

Although there is no consent on the true robustness of the Rasch model, not only because the observed violations vary greatly, the overall findings seem to be leaning towards a wider applicability than its strict specifications seem to indicate. Rasch-trees as a method for DIF detection on the other hand seem to be vulnerable to some variations of parameters, though not all. This is an anomaly which can not be explained with the current results.

Limitations

As with every simulation study, the results can only be applied to the considered scenarios. The work at hand is explorative, but decisions over which conditions were simulated had to be made which leads to the exclusion of many other possible conditions. Especially the terms under which DIF were simulated were very constrained. Categorical covariables with more than two categories as well as continuous covariables with more than one cut point were excluded. More complicated DIF, for example with a u-form, interactions between the variables or cancellation, were also not explored. This is unfortunate, because the built-in search for a cut point and the graphic presentation off Rasch-trees are their most unique asset. This means, it lends itself to situations where more complex DIF is to be expected. An exploration of truly continuously simulated DIF would surely be of interest as well to test the cut point estimation of Rasch-trees. The conditions with more than two subgroups suggest that the Rasch-trees could struggle with more complicated DIF under higher level models.

Basic parameters like test length and sample size were not altered as well. This is especially unfortunate, as accurate estimations in higher level IRT models require longer tests (Slinde & Linn, 1979). Both of these parameters were restrained in the current simulation study by the available computing power and time, but should be

expanded in future work. The current study only presented very simple influences DIF can take. This allowed the exploration of other interesting connections like scenarios, where DIF occurs in more than one parameter or the differentiation on how many items are influenced by DIF. Nevertheless, the results have to be interpreted with the limited conditions in mind.

Conclusion

To come back to the research question: Can the Rasch-tree method still detect DIF in data, which was simulated under higher level IRT models? It has to be acknowledged first that Rasch-trees were designed to detect DIF in the item easiness parameter. The current simulations deliberately broke this intended use and also manipulated DIF in other item parameters. Therefore, if DIF is detected, it is not identified correctly, but misplaced as DIF in the item easiness in the conditions were there is no DIF in the item easiness. This is certainly preferable to not detecting it at all. In general, the Power can reach desirable levels if the DIF value is high enough under the 2PL model. Detection was less likely for the new parameters of the 4PL model. The estimation accuracy of the item easiness parameter is highly disturbed under the 4PL model, regardless of DIF conditions, less so with the 2PL model. Here, it is more dependent on the specific condition. No condition resulted in drastically more subgroups. The Rasch-trees still seem to be able to detect the correct number of necessary cuts when DIF was simulated regarding one covariable. The Type-I-Error rate was mostly unaffected by the varying conditions. In this regard, Rasch-trees offer a possibility for detecting DIF in the item discrimination as well, less in the guessing or slipping error parameter. The consequences of using data with model misspecifications are generally worse under the 4PL model and lean towards missing DIF, which is present in the data. There is certainly no elevated risk of finding DIF, which is not present.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, *29*(1), 67–91.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society: Series B (Methodological)*, *32*(2), 283–301.

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*(1), 123–140.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), i–8.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459.

Costa Jr, P. T., & McCrae, R. R. (2008). *The revised neo personality inventory (neo-pi-r)*. Sage Publications, Inc.

Dinero, T. E., & Haertel, E. (1977). Applicability of the rasch model with varying item discriminations. *Applied Psychological Measurement*, *1*(4), 581–592.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the mantel–haenszel method. *Applied Measurement in Education*, *2*(3), 217–233.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the rasch model. *Applied Psychological Measurement*, *5*(2), 175–186.

Glas, C. A. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika*, *64*(3), 273–294.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing dif. *Psychometrika*, *61*(4), 647–677.

Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: an international journal*, *40*(10), 1679–1694.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rasch, G. (1966). An individualistic approach to item analysis. *Readings in mathematical social science*, 89–108.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models.

Applied Psychological Measurement, *10*(3), 303–317.

Slinde, J. A., & Linn, R. L. (1979). The rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement*, *3*(4), 437–452.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the rasch model. *Psychometrika*, *80*(2), 289–316.

Suárez-Falcón, J. C., & Glas, C. A. (2003). Evaluation of global testing procedures for item fit to the rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127–143.

van de Vijver, F. J. (1986). The robustness of rasch estimates. *Applied Psychological Measurement*, *10*(1), 45–57.

Appendix

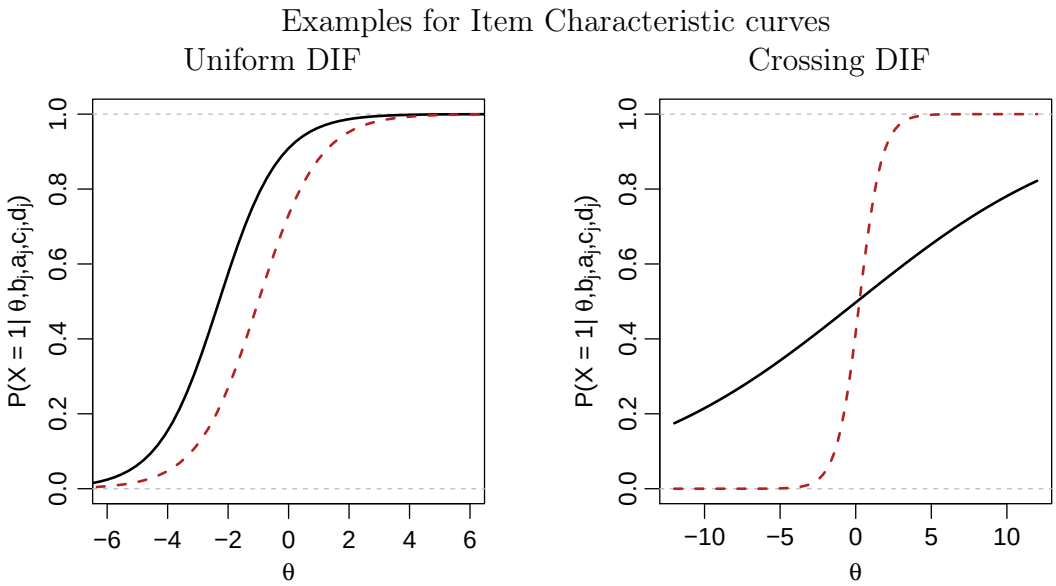

Figure 9. Exemplary Item characteristic curves for one of the simulated data sets with uniform or crossing DIF in a categorical covariable.

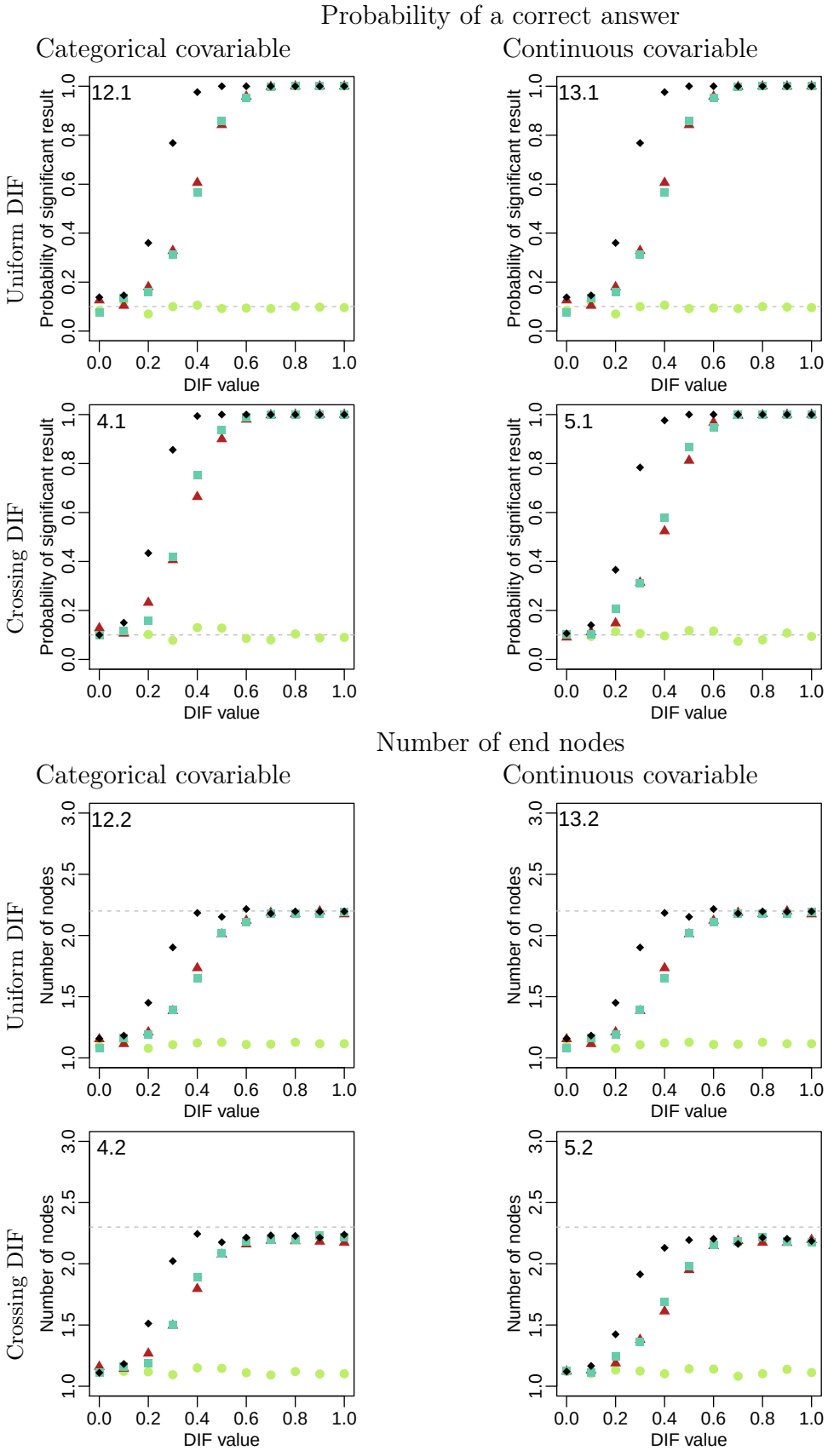

Figure 10. Probability of a correct answer and Number of end nodes in the Rasch-trees dependent on the scale of DIF and the number of affected items (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items affected) with crossing DIF or uniform DIF.

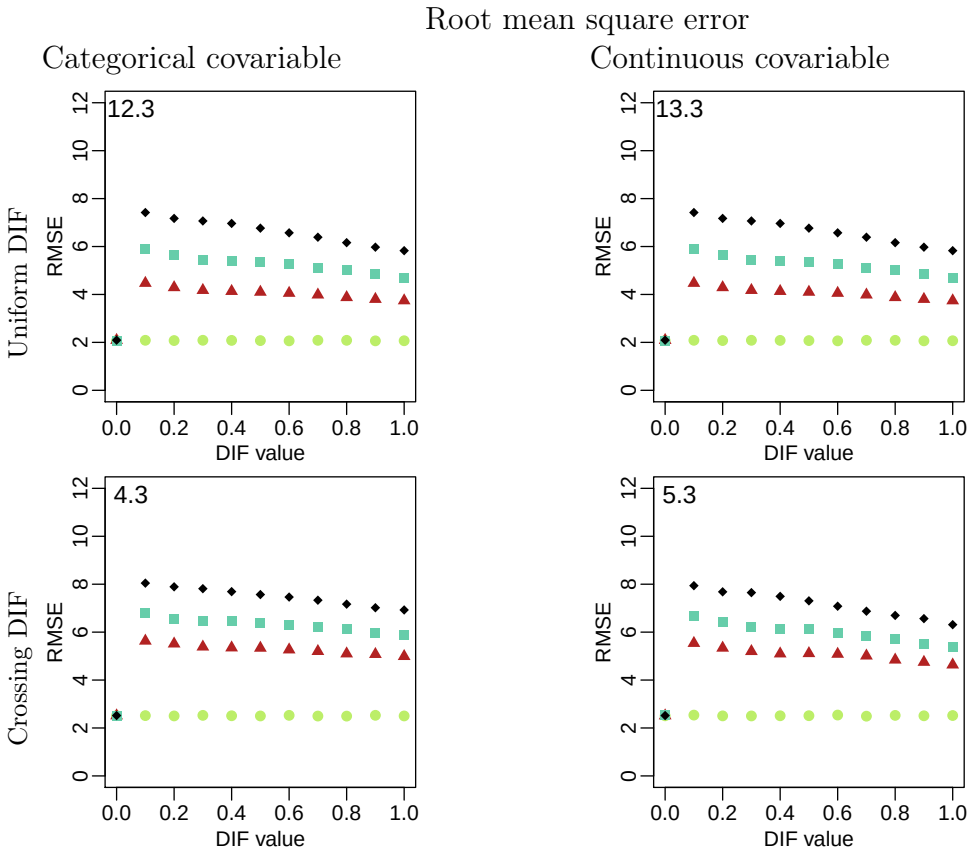

Figure 11. Root mean square error (RMSE) dependent on the value of DIF and the number of affected items (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items affected) with crossing DIF or uniform DIF.

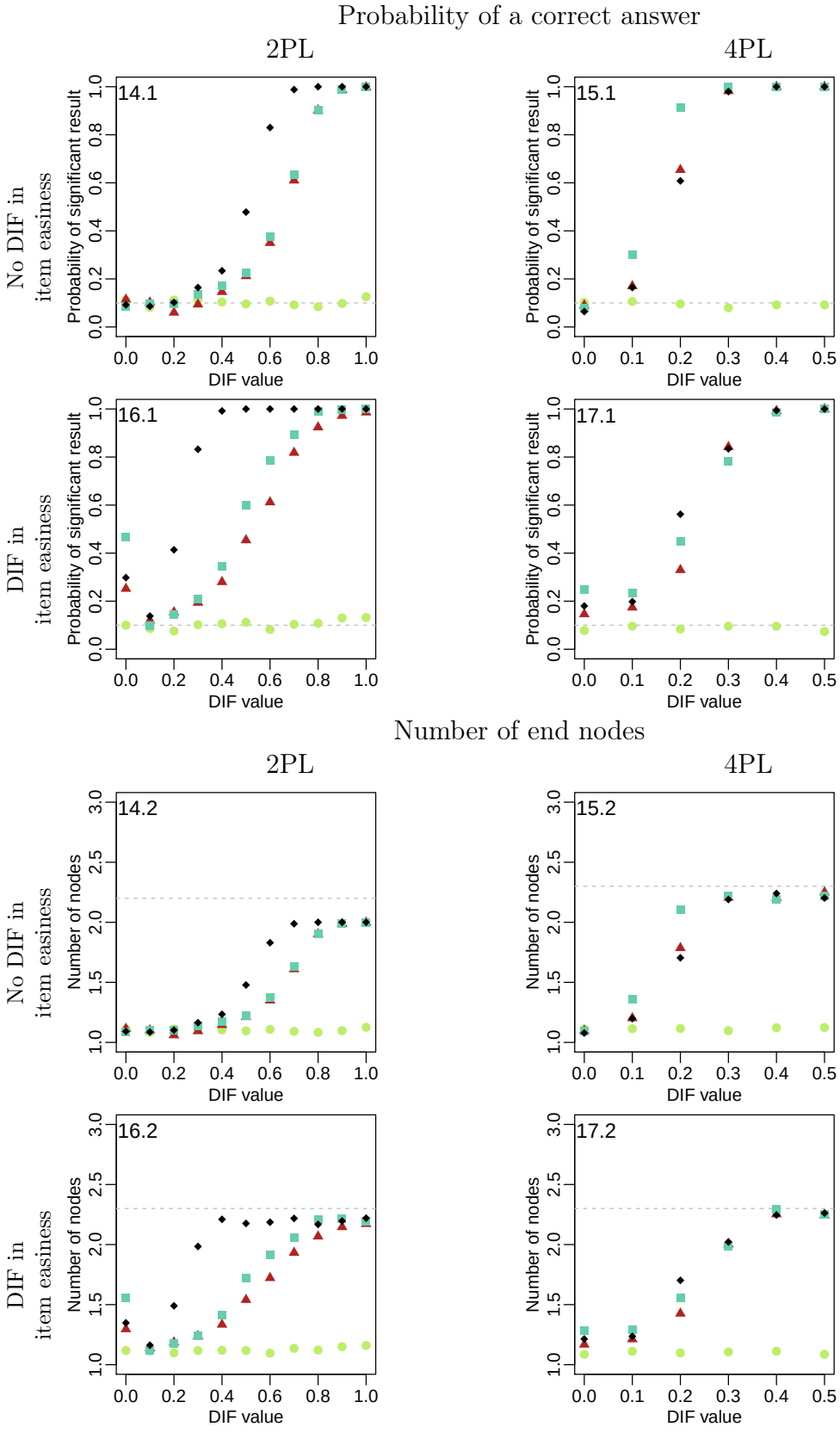

Figure 12. Probability of a correct answer and Number of end nodes in the Rasch-trees dependent on the scale of DIF and the number of affected items (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items affected) with DIF in both a categorical covariable and a continuous covariable.

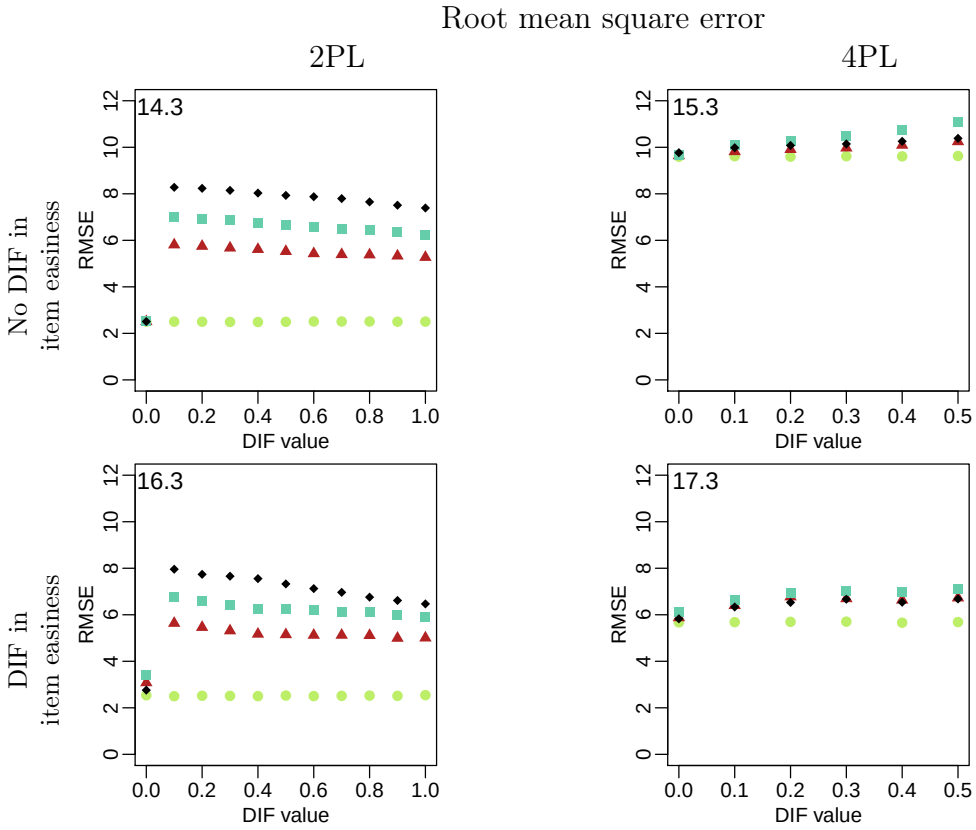

Figure 13. Root mean square error (RMSE) dependent on the scale of DIF and the number of affected items (green dots = 0 items, red triangles = 2 items, blue cubes = 5 items, black diamonds = 8 items affected) with DIF in both a categorical covariable and a continuous covariable.