One Model May Not Fit All

Subgroup Detection Using
Model-Based Recursive Partitioning

Marjolein Fokkema[1], **Mirka Henninger**[2,3] **and Carolin Strobl**[2]

[1]Leiden University, [2]University of Zurich, [3] University of Basel

One Model May Not Fit All: Subgroup Detection Using Model-Based Recursive Partitioning

Abstract

Model-based recursive partitioning (MOB, Zeileis, Hothorn, & Hornik, 2008) is a flexible framework for detecting subgroups of persons showing different effects in a wide range of parametric models. It provides a versatile tool for detecting and explaining heterogeneity in, for example, intervention studies. In this tutorial paper, we provide an introduction to the general MOB framework. In two specific case studies, we show how MOB-based methods can be used to detect and explain heterogeneity in two widely-used frameworks in educational studies: the generalized linear mixed model (GLMM) and item response theory (IRT). In the first case study, we show how GLMM trees (Fokkema, Smits, Zeileis, Hothorn, & Kelderman, 2018) can be used to detect subgroups with different parameters in mixed-effects models. We apply GLMM trees to longitudinal data from a study on the effects of the Head Start pre-school program, to identify subgroups of families where children show comparatively larger or smaller gains in performance. In a second case study, we show how Rasch trees (Strobl, Kopf, & Zeileis, 2015) can be used to detect subgroups with different item parameters in IRT models, i.e. differential item functioning (DIF). DIF should be investigated before using test results for group comparisons. We show how a recently developed stopping criterion (Henninger, Debelak, & Strobl, 2023) can be used to guide subgroup detection based on DIF effect sizes.

Introduction

Model-based recursive partitioning (MOB; Zeileis et al., 2008) allows for detecting subgroups of persons that differ in the parameters of a statistical model. To find the subgroups, MOB uses *recursive partitioning*, a technique that was first popularized by the classification and regression trees (CART; Breiman, Friedman, Olshen, & Stone, 1984) algorithm. In CART, the aim is to detect groups of persons, defined by specific (combinations of) covariate values, that differ in their mean on a response variable. Detecting such subgroups may be of interest in many studies in school and developmental psychology. For example, recursive partitioning methods have been used to identify adolescents at risk of committing crime (Fritsch, Haupt, Lösel, & Stemmler, 2019) and experiencing corporal punishment (Stemmier, Heine, & Wallner, 2019), and to identify predictors of math ability (Ding & Zhao, 2019).

Like CART, MOB also identifies groups of persons defined by (combinations of) covariate values, but the groups can differ in a wider range of parameters, instead of only the mean. For example, MOB can be used to identify children that show stronger or weaker effects of an intervention (Chung, Ansong, Brevard, & Chen, 2021), or stronger or weaker gains in academic skills over time (Fokkema & Zeileis, in press). The framework of MOB is very flexible and can be applied to a wide range of statistical models, such as linear and logistic regression (Kopf, Augustin, & Strobl, 2013; Zeileis et al., 2008) or models for paired-comparison data (Strobl, Wickelmaier, & Zeileis, 2011; Wiedermann, Frick, & Merkle, 2021). In this article, we will highlight two specific MOB methods that may be particularly relevant for research in school psychology: MOB for mixed-effects models (Fokkema et al., 2018) and MOB for Rasch measurement and Item Response Theory models (IRT, Henninger et al., 2023; Komboz, Strobl, & Zeileis, 2017; Strobl et al., 2015).

Mixed-effects models become relevant whenever data are collected in longitudinal or nested data structures. For example, when children are tested at several time points, observations at each timepoint are nested in children; when children from different classes from different schools participate in a study, children are nested in classes, which are in turn nested in schools. In the section MOB for Subgroup Detection in Mixed-Effects Models, we present an example application to illustrate how MOB can be used to detect subgroup-specific intervention effects, while taking into account the nested data structure. In this example, the focus is on detecting subgroups with different parameters of a *regression* model. We thus assume that the psychological test score(s) of interest have already been validated.

In the section MOB for Detecting DIF in Measurement Models, we go back in the research process to the point where a new psychological test has been administered to a validation sample. Here, we focus on detecting subgroups with different parameters of a *measurement* model. Validity assessment requires that we make sure that test results are comparable, for example between children of different genders. If certain items show different measurement parameters between groups, this may put certain groups at a relative disadvantage. These items are said to violate measurement invariance or exhibit differential item functioning (DIF). In order to test for DIF in the framework of Item Response Theory or Rasch modelling (Anthony, DiPerna, & Lei, 2016; Debelak, Strobl, & Zeigenfuse, 2022; Maller, 1997), the item parameters are compared between groups of persons. This can be done in a way that allows to detect DIF, while accounting for possible true group differences in ability. Traditional approaches for testing DIF require the groups to be pre-specified in order to test for DIF. In an application example we show how MOB flexibly allows to detect groups with different item parameters in a data-driven way.

In the following section, we first give a short introduction into the algorithm and statistical concepts behind MOB. Readers interested in learning more about its predecessor method, classification and regression trees, are referred to the introduction by Strobl, Malley, and Tutz (2009). Our aim here is to provide an informal introduction and to illustrate the relevance of these methods for studies in school psychology. For readers interested in the more formal details of MOB and the extensions presented here, relevant references are provided throughout the manuscript.

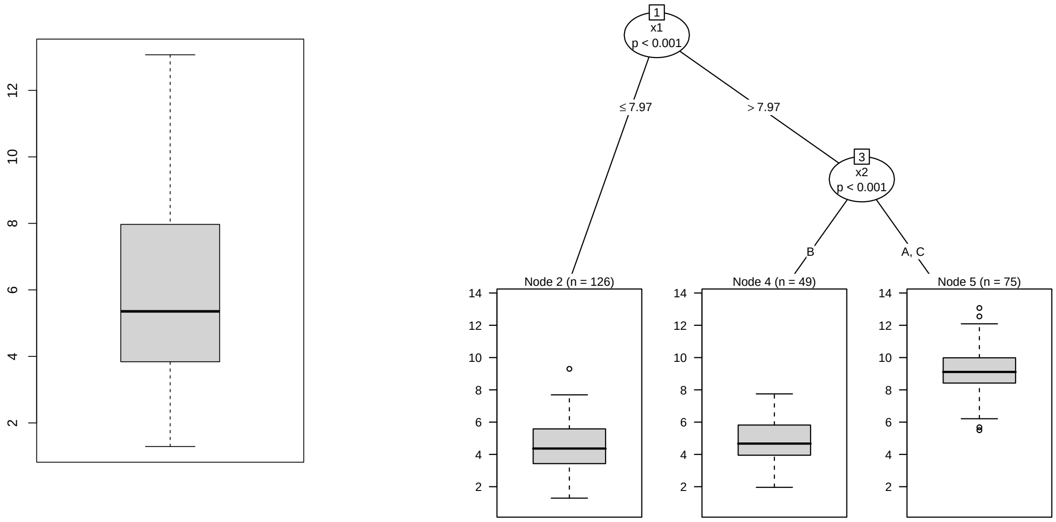
The MOB Algorithm

The main rationale of MOB is that one global model may not fit all observations in a dataset equally well. In many studies, additional covariates may be available. It may then be possible to uncover subgroups defined by these covariates, and obtain better-fitting models in each of those subgroups (Zeileis et al., 2008).

We illustrate this idea using a simple, simulated toy dataset, comprising 250 observations and four variables: A continuous response variable y (e.g., a total score for behavioral difficulties), and three covariates, x_1 (e.g., age), x_2 (e.g., a reading comprehension score) and x_3 (e.g., gender, with levels male, female and non-binary), as possible partitioning variables. To keep the example simple, we apply MOB to a very basic global model, which comprises only an intercept (or mean). It would also be possible to use, for example, a linear or logistic regression model as the global model. Figure 1, left, shows the distribution of y in a boxplot. Obviously, a global intercept or mean would not describe all observations equally well, because there is quite some unexplained variation around the sample mean of 5.91.

Figure 1

Left: Univariate distribution of the response variable. Right: Tree with group-specific distributions of the response variable in the terminal nodes.



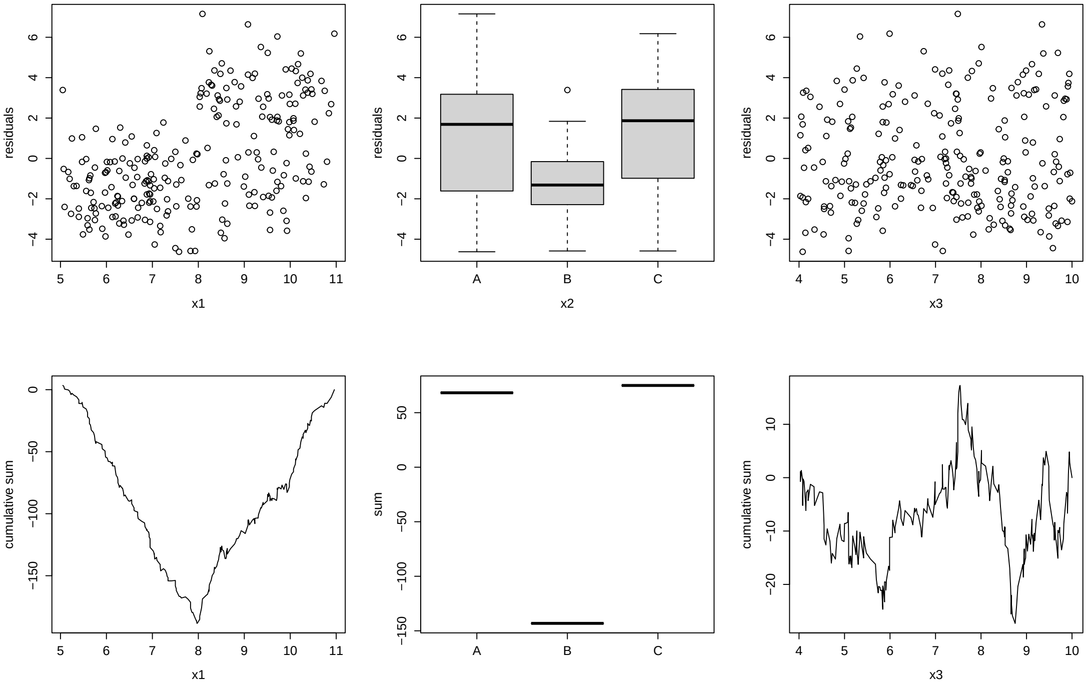
To detect possible subgroups with different values for the parameters, MOB cycles iteratively through the following steps:

1. The model parameters are first estimated jointly for all persons in the current node, starting with the root node containing the full sample.

2. Structural change in the model parameters is assessed with respect to each available

Figure 2

Top row: Residuals ordered by the values of the partitioning variables. Bottom row: Summed residuals (cumulative sums for continuous predictors, sums by level for categorical predictors).



covariate.

3. If there is significant structural change, the observations in the current node are split using the covariate associated with the strongest change.

4. Steps 1–3 are repeated recursively in each resulting node until there is no more significant structural change (or the groups become too small).

We applied MOB to the observations in the left panel of Figure 1, specifying x_1, x_2 and x_3 as potential partitioning variables. In Step 1, the model parameters (here: the intercept) are estimated in the root node, which contains all observations. In Step 2, the model's *scores* are computed. In case of an intercept-only model, these are simply the residuals. The scores are ordered with respect to the possible partitioning variables, as shown in the top row of Figure 2. The scores should randomly fluctuate around their mean of zero, but the top row of Figure 2 suggests systematic deviations: there seems to be an association with x_1 and x_2, and little or no association with x_3.

MOB captures systematic deviations in the scores (i.e., structural change) by computing cumulative sums. The cumulative sums are depicted in the bottom row of Figure 2. To decide in Step 3 whether to split the node, and if so, which variable to split on, a test statistic is computed based on the cumulative sums of each partitioning variable; full details on the computation are provided in (Zeileis et al., 2008). In the current example, the structural change tests revealed that x_1 was most strongly associated with instabilities in the intercept and it was thus selected for splitting in Step 3.

In Step 4, the procedure is repeated in each of the resulting subgroups. The resulting tree is shown in the right panel of Figure 1. In the right subgroup (node 3), additional significant instability was detected with respect to x_2, a categorical covariate. Again, the cutpoint for x_2 was selected so that the two resulting subgroups exhibit an as large as possible parameter difference. In the left subgroup (node 2), no further splits were created, because none of the three covariates were significantly associated with any further instability in this subgroup. The same held for nodes 4 and 5, so that the third covariate x_3 was never selected for splitting.

The p values resulting from the structural change tests of each split are depicted in the corresponding nodes. By default, the p values are Bonferroni corrected, to account for the fact that the tests are performed for all potential partitioning variables. After selecting x_1 for splitting, the cutpoint is selected so that the two resulting subgroups exhibit the strongest parameter differences, while the observations within the subgroups are as similar as possible.

The subgroup-specific distributions of the response variable are depicted in the terminal nodes at the bottom of the tree. The fact that the tree in Figure 1 displays more than one terminal node confirms that one global model for all observations does not appropriately capture the patterns in the data. The tree also shows shows that, out of the three covariates that were presented to the algorithm, only two were selected for splitting. This automatic variable selection is an important characteristic of the CART and MOB algorithms.

The means (intercepts) in the terminal nodes are 4.42, 4.78 and 9.17, respectively. While the means and distributions in nodes 2 and 4 appear very similar, the mean in node 5 is obviously higher. Substantively, this tree suggests an interaction effect: The effect of x_2 depends on the level of x_1 (vice versa). If x_1 were age and x_2 gender, this would indicate that gender differences only start to occur at a later age.

There are three further important characteristics of tree and MOB algorithms that we would like to mention here: The first characteristic is the type of variable and cutpoint selection an algorithm employs. Traditional classification and regression tree algorithms, like those of Breiman et al. (1984) and Quinlan (1993), performed variable and cutpoint selection in one step, which leads to an undesirable behavior called variable selection bias. That is, the traditional algorithms prefer variables offering more possible cutpoints in the selection process – regardless of their true information content. An algorithm that still has this problem is, for example, the `rpart` algorithm in R, based on the original CART algorithm by Breiman et al. (1984). More modern algorithms for classification and regression trees have solved this problem and offer unbiased variable selection, such as `QUEST` (Loh & Shih, 1997), which is available in SPSS, and `ctree` (Hothorn, Hornik, & Zeileis, 2006), which is available in `R` in packages `party` and `partykit` (Hothorn & Zeileis, 2015). The latter forms the basis for all MOB approaches presented in this paper. For more details on the statistical theory behind unbiased classification and regression trees and MOB, see Hothorn et al. (2006), Strobl et al. (2009) and /citetStroyKopf15.

The second characteristic relates to the way a tree or MOB algorithm stops splitting: Modern algorithms for classification and regression trees and MOB use a criterion of statistical significance to stop splitting. Once there are no more covariates that show a significant structural change in any node, splitting is halted. In this way, the algorithm selects only those partitioning variables that are relevant for distinguishing the groups (i.e., it performs automatic variable selection, as illustrated in the right panel of Figure 1). Moreover, the trees will not grow as large as possible, but splitting is stopped when no more significant structural change is detected. While traditional tree algorithms like those of Breiman et al. (1984) and Quinlan (1993) grew very large trees and then cut them back (so called *pruning*), modern tree and MOB algorithms employ significance tests as stopping criteria (and some can also use effect size measures, as we will see in Application Example 2). This allows for stopping tree growing as soon as no significant differences can be detected anymore. Other stopping criteria are based on the number of persons in the terminal nodes. These criteria ensure that the sample sizes in the terminal nodes are large enough to estimate the statistical model in each terminal node. For reasons of interpretability, users may specify even higher requirements for the sample sizes of terminal nodes.

The third characteristic of classification and regression trees as well as MOB is that the entire structure identified by the trees does not have to be pre-specified by the researcher in a confirmatory manner, but is learned from the data in an exploratory manner. This is a key feature of the MOB approach that makes it very flexible and sets it apart from purely parametric approaches, where only those main effects and interactions that are explicitly included in the specification of the model are considered. While there are phases in psychological and educational research where it is very important to specify hypotheses a-priori and test them in a confirmatory manner, in early stages of research exploratory methods are an important addition to the statistical toolbox for researchers. Still, an important challenge for the researcher remains: To specify the parametric model of interest, to specify the set of possible partitioning variables and to choose the settings of the MOB algorithm. The current paper aims to provide guidance.

Next, we discuss two specific methods that make use of the general MOB framework introduced above. The methods allow for partitioning two types of statistical models that are particularly relevant for school psychology research: Mixed-effects models for repeated measures or nested data structures and measurement models for validating psychological and educational tests.

MOB for Subgroup Detection in Mixed-Effects Models

Mixed-effects models contain two types of effects: Fixed and random effects. Fixed effects are typically used to capture population-averaged effects, while random effects are used to capture inter-individual variation around these fixed effects. In many studies, researchers are specifically interested in testing hypotheses relating to the population-averaged effects, while the random effects are included in the model to properly account for inter-individual variation, and correlations between observations within the same unit (Raudenbush & Bryk, 2002).

GLMM trees combine MOB and generalized linear mixed-effects models (GLMMs) and were introduced by Fokkema et al. (2018). While the 'standard' MOB trees (Zeileis et al., 2008) allow for subgroup detection in fixed-effects GLMs, GLMM trees additionally estimate and account for random effects and can thus be used for subgroup detection in mixed-effects regression models. The GLMM tree algorithm only targets structural change in the fixed-effects parameters. The random-effects parameters are specified as usual and assumed constant; that is, they are estimated using all observations in the dataset. The resulting subgroups will differ in their estimates for the fixed-effects coefficients only. As such, GLMM trees allow for detecting subgroups in multilevel models that differ with respect to any (set of) fixed-effects parameters of interest. For example, users may be interested in detecting subgroups with different means (Fokkema, Edbrooke-Childs, & Wolpert, 2021), but also differential effects of treatment (Fokkema et al., 2018) or differential growth over time (Fokkema & Zeileis, in press), to name but a few examples.

The `R` functions for fitting GLMM trees are `lmertree` and `glmertree`. The former assumes a normally distributed response, while the latter supports, for example, binomial or count responses. The following section will illustrate the usage of `lmertree`. Mathematical and computational details of the GLMM tree algorithm are described in Fokkema et al. (2018) and Fokkema and Zeileis (in press).

Application Example 1: Detecting Differential Effects of Head Start

Dataset

To illustrate the use of GLMM trees, we analyze a dataset from Deming (2009), who evaluated long-term benefits of participation in the Head Start program. Head Start is a federally funded nationwide pre-school program for children from low-income families in the United States. Participation in Head Start takes place from ages 3 through 5. Deming (2009) compared performance of siblings who differed in their participation in the program using data from the 1979 cohort of the National Longitudinal Survey of Youth (Bureau of Labor Statistics, U.S. Department of Labor, 2019).

The sample consists of 273 families with at least two siblings, where at least one sibling participated in Head Start and at least one sibling did not participate in Head Start or any other pre-school program. Data from children who participated in another pre-school program were excluded. The family structure allows siblings who did not participate in Head Start to serve as a natural control to assess the effects of participating in Head Start. The outcome variable comprises repeated assessments on the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1981), for which we model trajectories over time. On average, there were 2.14 PPVT scores per child, for 71% of the children there were > 1 PPVT scores available.

Our dataset contains five family characteristics: Mother's score on the Armed Forced Qualification Test (AFTQ), adjusted for age; family income (averaged over the years for which data was available); race (Black, Hispanic or White); mother's years of completed education; mother's height (computed as feet x 100 + inches). Note that the latter variable is one that should be completely irrelevant for predicting performance on a vocabulary test; it is included here to illustrate that the GLMM tree algorithm can fruitfully distinguish signal from noise variables. The dataset analyzed here only comprises data from families and children for whom complete data was available.

We load the data and inspect the first rows:

```
load("HS_dat.Rda")
head(HS_dat, 3)
```

```
       AFTQ     Race   Income Mom_height Mom_edu_yrs ChildID MotherID Program
1  3.478122 Hispanic 37731.07        502          12   20502      205      HS
2  3.478122 Hispanic 37731.07        502          12   20501      205    None
3 15.964368    Black 16119.13        504          10   22403      224    None
  PPVT Age Age_orig
1   18   4        4
2   48   7        7
3   69   7        7
```

```
hist(tapply(HS_dat$Mom_height, HS_dat$MotherID, mean))
```

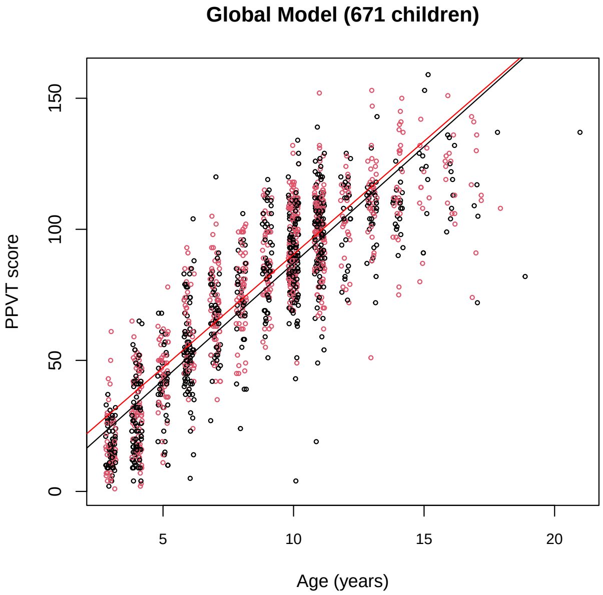
We first inspect the complete dataset by plotting PPVT scores against age, separated by program participation: None versus Head Start. To show the global effect of age and Head Start participation, we first fit a mixed-effects model comprising their main and interaction effects, using package lme4. This would also be the model fitted in the root node of the GLMM tree:

```
library("lme4")
lmm <- lmer(PPVT ~ Program*Age + (1|MotherID/ChildID), data = HS_dat)
```

We specified the fixed main and interaction effects of ‘Program‘ and ‘Age‘ as ‘Program*Age‘. The random effects were specified as ‘(1|MotherID/ChildID)‘. The 1 specifies that a random intercept should be estimated and MotherID/ChildID specifies that every child and every mother should obtain a random intercept, so as to account for correlations between repeated assessments of the same child, and correlations between children of the same mother. The forward slash / indicates that the children are ’nested’ within mothers; that is, that there are multiple children per mother, but only one mother per child.

The results are presented in Figure 3, which shows that children participating in Head Start show slightly higher performance than their non-participating siblings and that this difference persists but slightly diminishes over time. This result agrees with the findings of Deming (2009). Figure 3 also shows that the effect of age on performance is very strong, which is typical for young children growing up to adults, while the effect of intervention (Head Start) is small in comparison. Note that the R code for creating the figure is omitted here, because we want to focus on fitting and interpreting GLMM trees. Code for exact replication of the results presented here is provided in the Supplementary Materials.

Figure 3



Fitting a GLMM tree

We now test whether the intercepts and slopes of the two regression lines differ as a function of the partitioning variables, using function `lmertree` from R package **glmertree**:

```
library("glmertree")
HS_tree <- lmertree(PPVT ~ Program*Age | (1|MotherID/ChildID) | AFTQ + Race +
                                         Income + Mom_edu_yrs + Mom_height,
                                       data = HS_dat, cluster = MotherID, minsize = 250)
```

With the first argument, we specified the model `formula`, which has three parts separated by vertical bars: The left part (`PPVT ~ Program*Age`) specifies the response variable, followed by a tilde (`~`) and the fixed-effects predictors of relevance. The middle part (`1|MotherID/ChildID`) specified the random effects, which are identical to the random effect in the earlier mixed-effects model. The right part (`AFTQ + Race + Income + Mom_edu_yrs + Mom_height`) specified the partitioning variables: covariates that may possibly affect the values of the fixed-effects parameters.

With the second argument, we specified the dataset which contain the variables. With the `cluster` argument, we account for the level at which the partitioning variables ae measured. In this example, the partitioning variables are measured at the mother (or family) level, which we indicated by specifying `cluster = MotherID`. As a result, the parameter stability tests will employ so-called clustered covariances (Zeileis, Köll, & Graham, 2020). Not specifying the `cluster` argument would result in use of the default observation-level covariances in the parameter stability tests, inflating type-I error rates when partitioning variables are measured on a higher level (Fokkema & Zeileis, in press).

To aid interpretation, we would like to retain large enough subgroups. We therefore specified the `minsize` argument, so that splits will only be implemented if the resulting nodes contain at least 250 observations. With a total sample size of 1433, even small differences in the effects of age and Head Start participation between subgroups may become significant. The average number of measurements per family was 5.6. A minimum node size of 250 ensured that the estimated effects of age and Head Start participation in the terminal nodes would at least be based on data from about 50 families.

Interpreting a GLMM tree

Next, we plot the tree. With multiple fixed-effects predictors of interest, the default plots may become too crowded or difficult to interpret. We therefore specify `type = "simple"` to facilitate interpretation, and using the `nodesize_level` argument, we specified that the sample size printed above every terminal node should count the number of children, instead of the number of individual observations, which would be printed by default:

```
plot(HS_tree, type = "simple", nodesize_level = 2)
```

Table 1

Node-specific predicted PPVT scores at different ages.

Node	Program	PPVT at age 6	PPVT at age 12
2	None	45.70	95.63
2	HS	49.48	98.02
5	None	50.36	107.43
5	HS	53.74	104.98
6	None	53.33	107.82
6	HS	59.99	115.70
7	None	59.64	117.13
7	HS	63.56	124.09

extit{Note.} For computing predictions, random effects were assumed zero. HS = Head Start; PPVT = Peabody Picture Vocabulary Test.

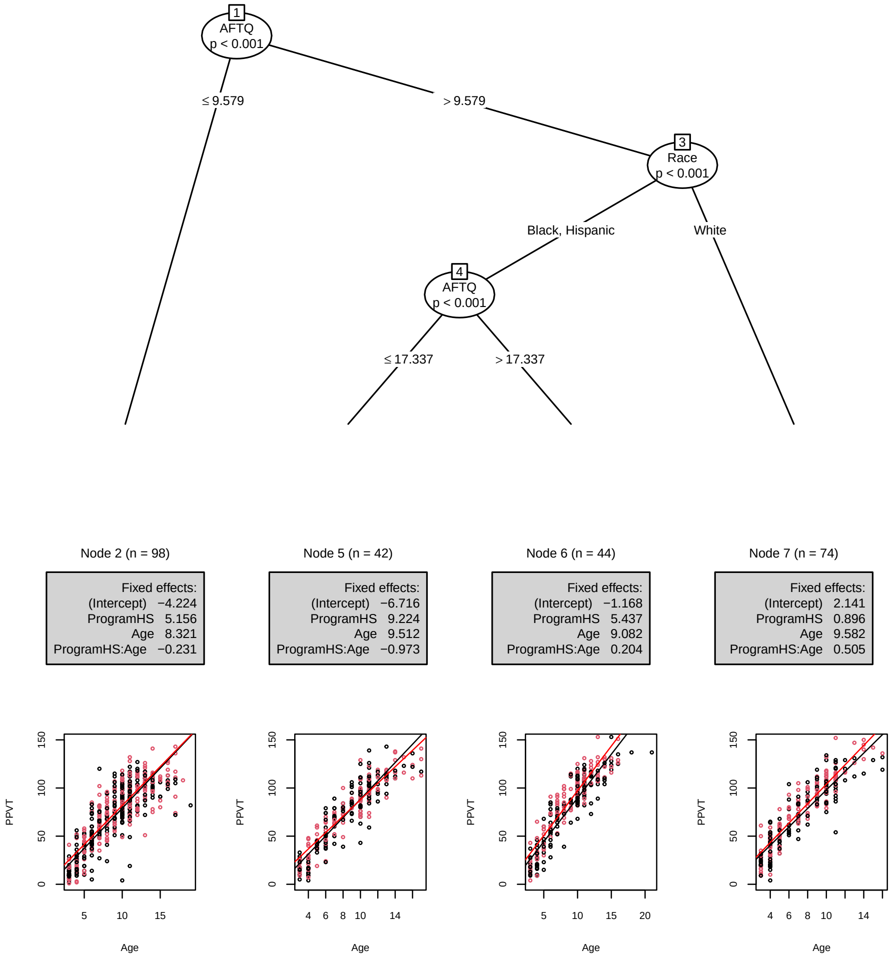
The resulting tree is presented in Figure 4. Below each terminal node, we plotted the observations and the two regression curves given by the coefficients in that node. The first split was made based on the AFTQ variable, which represents the mother's score on the Armed Forced Qualification Test, adjusted for the age at which they completed the test. The group with higher mother's AFTQ scores is further split based on race. The Black and Hispanic group is further split based on mother's AFTQ score.

To aid interpretation of the coefficients in Figure 4, Table 1 provides predicted PPVT scores for each of the groups and programs at ages 6 and 12. All nodes show a modest benefit of Head Start participation at age 6, about 3-4 points on the PPVT. This benefit increases over time for the group with higher mother's AFTQ scores (nodes 6 and 7). The benefit decreases over time for children in the group with lowest mother's AFTQ scores (node 2). For Black and Hispanic children with intermediate mother's AFTQ scores (node 5), the benefit at age 6 even shifts to a disadvantage of Head Start participation at age 12.

These results partly correspond to the findings of Deming (2009), who report a relative benefit of Head Start participation of about 0.15 standard deviations at age 5-6, which fades out over time to about 0.05 standard deviations at ages 11-14. Fade out was reported to be particularly strong for African-American and very disadvantaged children. In the current sample, the standard deviation of PVVT scores was 34.35, and thus the

Figure 4

GLMM tree with trajectories of PPVT scores



Note: Red lines depict average node-specific trajectories for siblings participating in Head Start; black lines depict average node-specific trajectories for siblings not participating in Head Start.

predictions for age 6 in Table 1 roughly correspond to the effect found by Deming (2009). However, we found a different pattern of fade-out.

It should be noted that MOB and GLMM trees are exploratory techniques: The subgroup structures are detected from the data. Such data-driven procedures cannot provide valid standard errors and significance tests to evaluate significance of the observed subgroup differences. To that end, the subgroup structure that was found should be used in a confirmatory analysis on a new sample, that was not used to fit the tree. This would allow to test the hypothesis, either using frequentist or Bayesian approaches, of whether the parameters of interest really differ between the detected subgroups.

If additional data is not available or cannot be collected, another possibility would be to split a single dataset into two parts before the analyses. Then one part of the dataset can be used for exploration (i.e., detecting the subgroups), and the other part can be used to test the parameter differences between the subgroups.

MOB for Detecting DIF in Measurement Models

Validity assessment of scores on psychological or educational tests requires researchers to assess whether the same construct is measured in the same way for different groups. In the framework of IRT and Rasch measurement, test items are typically investigated with respect to item misfit, multidimensionality and other violations of the measurement model (cf., for example Debelak et al., 2022, for an introduction). A particularly crucial assumption for the comparability of test scores between groups is measurement invariance. Items that violate measurement invariance by showing different measurement properties for different groups of participants exhibit differential item functioning (DIF).

MOB can be used to detect DIF by means of testing whether the item parameters of the measurement model exhibit significant instability with respect to (combinations of) covariates. For the Rasch measurement model, the `R` function for conducting the MOB analysis is the `raschtree` function from package `psychotree` (Strobl et al., 2015). The usage of this function will be illustrated in the next section (Application Example 2). We will see that, just like for GLMM trees, we need to specify which variables are part of the parametric model. In the case of the Rasch model, this will be the test items. Moreover, we need to specify which covariates are made available to the MOB algorithm for selecting relevant splitting variables and cutpoints.

If one joint Rasch model holds for the entire sample, that is, if there is no DIF, a Rasch tree should show no splits. However, in certain settings in educational research, such as large scale assessments, very large sample sizes are available for testing for DIF. Large sample sizes are good for detecting even small effects or model violations with high statistical power. The same holds for the statistical tests used for detecting parameter change in the MOB algorithm, so that in larger samples even very small parameter differences can be detected. However, in DIF detection this may mean that even very small DIF effects, that in practice can be considered ignorable, will be detected if only the sample is big enough. Therefore, an extension of Rasch trees has been suggested by Henninger et al. (2023) based on the Mantel-Haenszel effect size measure for DIF. Holland and Thayer (1985) have suggested an intuitive classification of DIF effect sizes based in the Mantel-Haenszel statistic, that is being widely used in educational testing. In this classification, category A stands for negligible DIF (small effect size or not statistically significant), B for medium DIF (neither A nor C), and C for large DIF (large effect size and statistically significant). Henninger et al. (2023) have incorporated this classification as an additional stopping criterion for Rasch trees, so that the user can decide, for example, that a split should only be conducted if the detected DIF is of category B or C, while negligible DIF of category A should be ignored.

Together with a purification step (see Henninger et al., 2023, for details), the Mantel-Haenszel classification can also be used for graphically highlighting those items that show DIF with respect to certain groups of persons. This can help generate hypotheses about the sources of DIF, and can also aid in deciding how to proceed with the DIF items. For example, items that show DIF between different language groups may often be improved by means of making sure that in all translations the meaning is as similar as possible, and/or that the words employed in the translations for the different languages are equally frequently used. In other situations, sources of DIF might be harder to eliminate, so that often DIF items may be excluded from a test. Either way, the measurement model needs to be refitted and its assumptions checked again after the final set of items has been decided upon and, in the case of modified items, administered again.

Another important aspect to keep in mind is that DIF can be caused by one or more items measuring a secondary dimension in addition to the dimension that is intended to be measured by the test. This can be the case, for example, for an instrument intended to measure math aptitude, that contains both pure algebra problems as well as story problems. Students whose native language is not the same as the test language can have a disadvantage in answering the story problems, for example when these contain seldomly used words. These items may then show DIF between native and non-native speakers. When encountering this, the test developers will have to decide whether the items with story problems should be excluded from the test, whether they can be improved, for example by using more frequently used words, or whether the test should be considered two- rather than one-dimensional. For a discussion of the connection between DIF and multidimensionality, see also Ackerman (1992) and Strobl, Kopf, Kohler, von Oertzen, and Zeileis (2021).

Application Example 2: Detecting DIF in a General Knowledge Quiz

Dataset

We will illustrate the usage of Rasch trees and the Mantel-Haenszel effect size measure by means of a dataset from Trepte and Verbeet (2010). An online general knowledge quiz was conducted by the German weekly news magazine SPIEGEL. Below we will use the abbreviation `SPISA` for the name of the data set, because the quiz was termed “students’ PISA” by the magazine, but it is not related to the “real” PISA study by the OECD. The data set contains the quiz results as well as sociodemographic data. In the following we will use only the responses to the nine items of the *culture* scale (i.e., items 28 through 36) from test booklet 20 and only the 9769 complete cases of university students for illustration. The wording of the items is provided in the Appendix.

Three possible partitioning variables are considered: `Gender` (in this data set coded only as male, female, or missing), `Age` (continuous), and `Area` (Area of study: Language & Culture, Law & Economics, Medicine & Health, Engineering, Sciences, Pharmacy, Geography, Agriculture & Nutrition, or Sports). The nine items from the culture scale are scored with 0 (incorrect answer) or 1 (correct answer). The quiz data have been prepared in a format where the nine items of the culture scale are not stored as individual columns of the data set, but the complete matrix of item responses is stored as a single variable (see Strobl, Schneider, Kopf, & Zeileis, 2021) under the name `culture`. This means that we can access the item responses for all items jointly using the name `culture` in the following `R` commands. The first few rows of the data set look like this:

```
load("dat_SPISA.Rda")
head(dat_SPISA, 3)
```

```
    Age Gender                    Area culture.i28 culture.i29 culture.i30
90   21   Male       Law and Economics           1           1           0
163  25   Male          no Information           1           1           1
402  22   Male Agriculture & Nutrition           0           1           0
    culture.i31 culture.i32 culture.i33 culture.i34 culture.i35 culture.i36
90            1           0           1           1           1           1
163           1           0           1           1           0           1
402           1           1           0           1           1           0
```

Fitting a Rasch tree

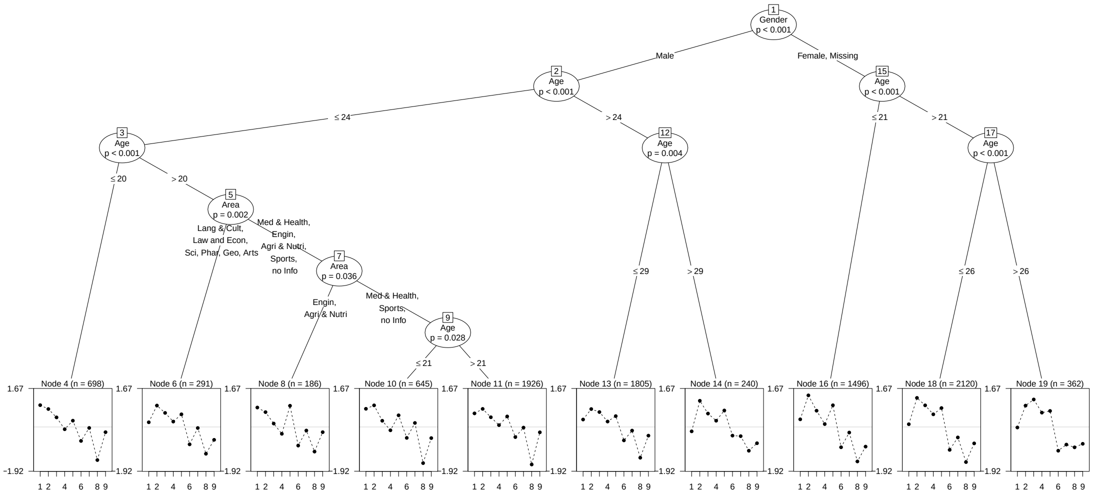
The results in Trepte and Verbeet (2010) indicate that the Rasch model is appropriate for the individual scales of the general knowledge quiz, so that we can continue to fit Rasch trees by means of function `raschtree` from the R package `psychotree`. Note that model-based recursive partitioning is also available in `psychotree` for other IRT models.

The model which we would like to test for parameter stability in this example, the Rasch model, is based only on the matrix of item responses `culture`. It does not contain predictor variables of its own. Therefore, the formula syntax for the `raschtree` function is simpler than that for the `lmertree` function in Application Example 1: On the left hand side of the `~` symbol, we provide the item response matrix `culture` for the Rasch model. On the right hand side of the `~` symbol, we provide the possible partitioning variables `Gender`, `Age`, and `Area`.

```
library("psychotree")
Raschtree_culture <- raschtree(culture ~  Gender + Age + Area,
                              data = dat_SPISA)
```

Figure 5

Rasch tree fitted to the SPISA quiz items using the default stopping criterion based on statistical significance.



Once the Rasch tree has been fitted, we can plot it using the standard `plot` function:

```
plot(Raschtree_culture)
```

The resulting tree is depicted in Figure 5. The item parameters for the nine culture items are displayed in the terminal nodes of the tree. However, it is hard to interpret the tree structure because such a large number of splits were created. Moreover, it is hard to tell whether all splits are due to substantial amounts of DIF, or whether some of the splits are due to small DIF effects that only became statistically significant due to the large sample size.

To assess whether DIF effects are substantial or negligible, it is possible to use the Mantel-Haenszel DIF effect size measure as an additional stopping criterion. It has three categories (A: negligible, B: moderate, C: large). If none of the items shows DIF in category B or C*, the tree is stopped from growing further. At the moment, the R functions for the Mantel-Haenszel stopping criterion and additional visualizations based on the Mantel-Haenszel effect size are available from GitHub. They will be made available as part of the `psychotree` package in the future.

*This is the default setting. It can also be changed so that splitting is stopped if no item shows DIF in category C.

In order to install the functions from the author's GitHub page and make them available in the current R session, use the following two commands:

```
library(devtools)
install_github("mirka-henninger/raschtreeMH")
library("raschtreeMH")
```

We can use a syntax very similar to that of the original raschtree function, but now we also specify the additional argument stopfun. We want to use the Mantel-Haenszel stopping function (stopfun_mantelhaenszel) together with iterative purification. It is also possible to provide other, user-defined stopping functions (see Henninger et al., 2023, for details).

For the purification strategy, there are three options: none, 2step and iterative. Purification means that items which have already been diagnosed with DIF are taken out of the sum score, which is used as the matching criterion in the final DIF test. It is highly recommended to use a purification strategy, because otherwise false positives (i.e., artificial DIF) may occur (cf., e.g., Debelak et al., 2022; Henninger et al., 2023; Kopf, Zeileis, & Strobl, 2015, and the references therein). In two-step purification, a new sum score without the items that were diagnosed with DIF in step 1 is computed for the final DIF analysis in step 2. When using iterative purification, this process is repeated until two runs yield the same DIF items or until the maximum number of iterations is reached. The final set of DIF-free items resulting from the purification will also be used for aligning the item parameters for each group comparison, which we will discuss and display below.

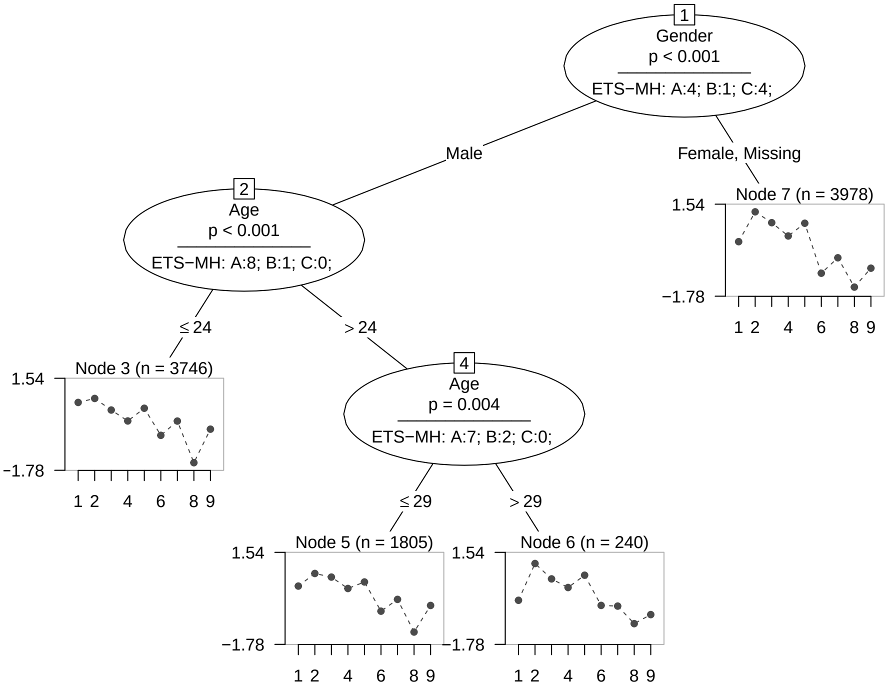
We fit the Raschtree using the Mantel-Haenszel stopping criterion as follows:

```
Raschtree_MH_culture <- raschtree(culture ~  Gender + Age + Area,
                                  data = dat_SPISA,
                                  stopfun= stopfun_mantelhaenszel(
                                    purification = "iterative"))
```

For technical reasons, the information about the Mantel-Haenszel effect size and classification is not saved in the tree object itself, but has to be added afterwards using the add_mantelhaenszel function.

Figure 6

Rasch tree fitted to the SPISA quiz items based on using Mantel-Haenszel DIF effect size measure as the stopping criterion.



```
Raschtree_MH_culture <- add_mantelhaenszel(Raschtree_MH_culture,
                                           purification = "iterative")
```

Interpreting a Rasch tree

After we have added this information to the Rasch tree, we can now visualize it with additional features:

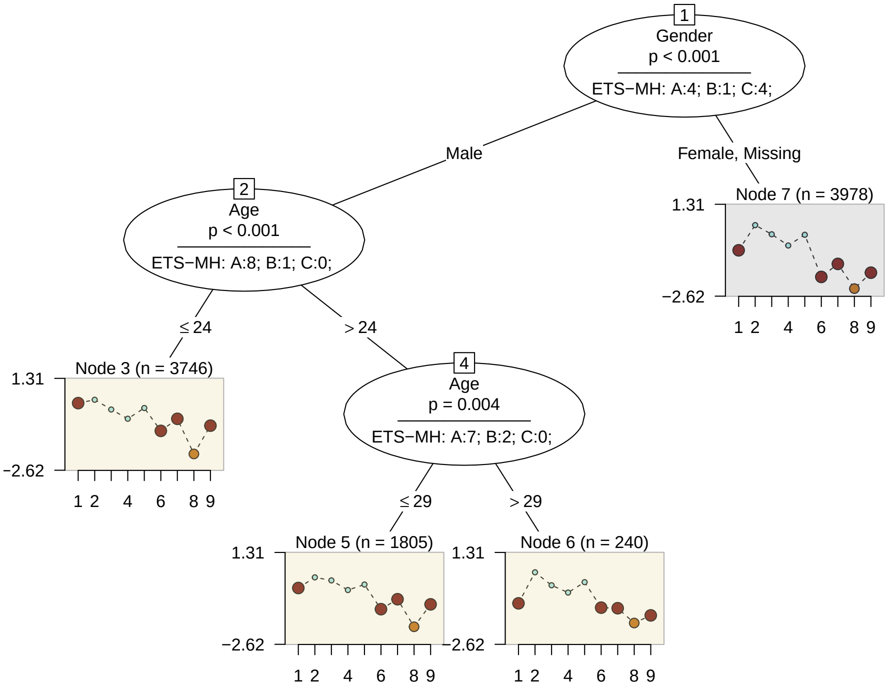
```
plot(Raschtree_MH_culture)
```

The result is presented in Figure 6, which reveals that the Rasch tree is much more concise with a lower number of splits. In addition, in each node the number of items classified as A, B, or C is displayed. For instance, we can see that in node 1 one item is classified in category B (moderate DIF), and four items are classified in category C (large DIF).

We can also color the DIF items in the terminal node profiles according to a partic-

Figure 7

Rasch tree fitted to the SPISA quiz items based on using Mantel-Haenszel DIF effect size measure as the stopping criterion with additional visualization features.



ular split in the tree. We will illustrate this with two examples: First, we can see that for node 1 (split in the covariate `Gender`), items 1, 6, 7, 8, and 9 show DIF in categories B or C. As displayed in the Appendix, the content of these items was:

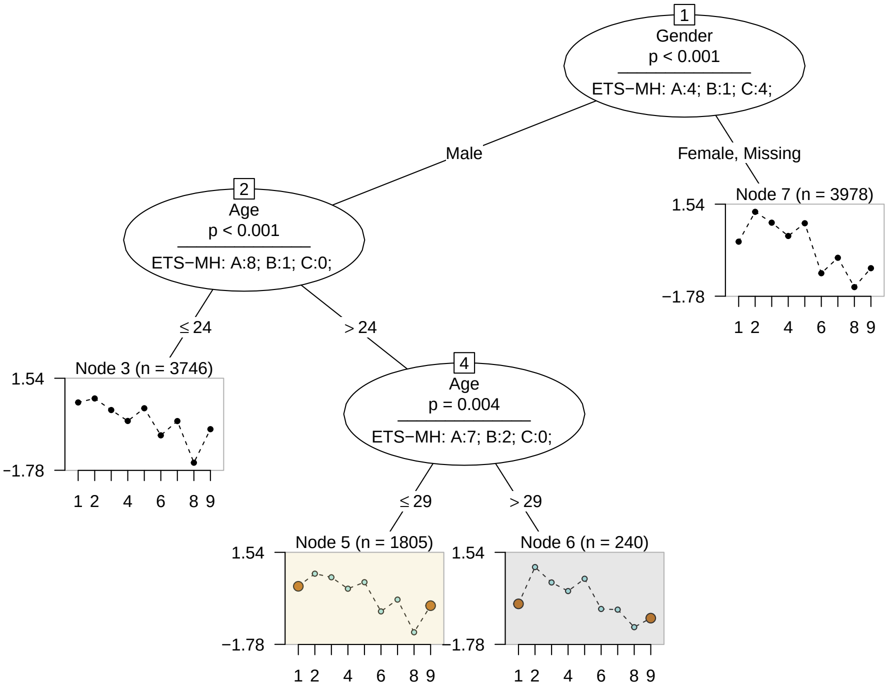
- 1: Painting by Andy Warhol
- 6: Novel by Daniel Kehlmann: Die Vermessung der Welt
- 7: City of the Buddenbrooks: Lübeck
- 8: City with building: Paris
- 9: Opera not by Mozart: Aida

```
plot(Raschtree_MH_culture, color_by_node = 1)
```

These items are easier or less difficult to answer for participants with female or missing gender (node 7), than for males (nodes 3, 5 and 6). Alternatively, we can also color the DIF items in the terminal node profiles according to a later split in the tree:

Figure 8

Rasch tree fitted to the SPISA quiz items based on using Mantel-Haenszel DIF effect size measure as the stopping criterion with additional visualization features.



```
plot(Raschtree_MH_culture, color_by_node = 4)
```

The result is presented in Figure 8. For node 4 (split in the covariate `Age`), items 1 (Painting by Andy Warhol) and 9 (Opera not by Mozart: Aida) show DIF and are more difficult to answer for younger male respondents ($24 < \text{age} \leq 29$), and easier or less difficult to answer for older male respondents ($\text{age} > 29$).

Detailed information about the Mantel-Haenszel criterion for each item in each node can also be accessed from the Rasch tree object using `$info$mantelhaenszel`. It contains the DIF classification and the value of the Mantel-Haenszel effect size measure for each item in each node, as well as the type of purification that was employed in each node, and in the case of iterative purification how many iterations were conducted.

```
Raschtree_MH_culture$info$mantelhaenszel
```

```
$classification
      node4 node2 node1
```

```
item1 "B"   "B"   "C"
item2 "A"   "A"   "A"
item3 "A"   "A"   "A"
item4 "A"   "A"   "A"
item5 "A"   "A"   "A"
item6 "A"   "A"   "C"
item7 "A"   "A"   "C"
item8 "A"   "A"   "B"
item9 "B"   "A"   "C"

$mantelhaenszel
            node4       node2        node1
item1 -1.53377668 -1.03559992 -1.802977840
item2  0.72945873 -0.14242027  0.377084860
item3 -0.35893502  0.51100000  0.127575191
item4 -0.05897255  0.40879442 -0.007385896
item5  0.41615400 -0.08996266  0.128463797
item6  0.42418247 -0.16529252 -1.875191790
item7 -0.87987154 -0.57949560 -1.620265527
item8  0.69240041  0.54735808 -1.083093749
item9 -1.00725807 -0.39502640 -1.798520587

$purification
      node4       node2       node1
"iterative" "iterative" "iterative"

$purification_counter
node4 node2 node1
    1     1     2
```

The displayed information again shows that some items show only negligible DIF (category A) in all nodes, while other items show medium (category B) or large (category C) DIF in some or all nodes. The values of the Mantel-Haenszel effect size measure are negative when the item difficulty is lower in the right than in the left child node. For example, as we already observed in Figure 7, items 1, 6, 7, 8, and 9 have medium or large DIF (DIF categories B and C) and are easier to answer for participants with female or missing gender (node 7).

Discussion

In this tutorial paper we have outlined the rationale of the MOB algorithm and how it can be used for detecting parameter differences in mixed-effects and Rasch models. The main advantage of MOB is that it can detect groups of persons with different model parameters in a data-driven manner. This makes it more flexible for detecting differences that were not hypothesized by the researcher. For example, it is often the case that obvious sources of DIF have already been avoided by the content experts during item creation. Any remaining DIF is unexpected. Therefore, DIF detection approaches that rely on the researcher to correctly specify the exact groups of persons that exhibit DIF can miss DIF, because it may be associated with other (combinations of) covariates than the ones investigated. In this sense, the more exploratory approach of Rasch trees has higher statistical power to detect DIF in previously unknown groups (Strobl et al., 2015). The same argument holds for parameter differences in mixed-effects models, where the substantial hypotheses formulated a priori are typically about global intervention effects or global effects of time, and not about possible subgroup-specific differences. GLMM trees allow for identifying possible interactions the predictor of interest may be involved in.

However, researchers should be aware that a covariate-based approach like MOB will only be able to detect parameter differences if the relevant covariates were recorded and supplied to the algorithm. If no (or few) potentially relevant covariates are available, an alternative framework for detecting previously unknown groups of persons with different model parameters is mixture modelling (see, e.g., De Ayala & Santiago, 2017; Frick, Strobl, & Zeileis, 2015, in the context of Rasch modelling). Mixture modelling aims at identifying latent classes of persons with different properties. It can also be combined with observed covariates. A comparison of MOB and mixture modelling is given by Frick, Strobl, and Zeiles (2014).

Another important caveat is that those covariates that are selected for splitting by the MOB algorithm are not necessarily causal for the observed parameter differences. They might also be proxies for other, unobserved covariates. For example, if DIF in the item parameters of a test on reading ability is found between different cities or districts, a variety of factors could be causally driving these effects, such as different compositions of native languages or sociodemographics. On the other hand, MOB can also be useful in evaluating causal effects, as it can be used to test for possible confounding (van Wie, Li, & Wiedermann, 2019).

As with any exploratory method, MOB should be considered as a tool for generating hypotheses and not as a confirmatory technique. While the MOB algorithm, as it is implemented in the R packages presented in this tutorial, has been constructed with care and uses statistical significance tests (and in the case of Mantel-Haenszel trees also effect sizes) to avoid spurious splits, for confirmatory hypothesis testing it is recommended to use a new dataset, not used for fitting the tree.

For example, the results of the GLMM tree analysis (Application Example 1) could be validated using later cohorts from the same study. On data from later cohorts, a confirmatory mixed-effects model could be fitted, similar to the one depicted in Figure 3, but with additional binary indicators distinguishing the subgroups. The model could be specified to include main effects as well as interactions with time and treatment (i.e., Head Start versus no Head Start). This refitting on new data will yield valid standard errors, and thus provide hypothesis tests on subgroup differences. Alternatively, if the interest of the analyses was primarily to identify moderators of the effect of Head Start, the model fitted on a later cohort could comprise main and interactions effects of time, Head Start, Mother's AFTQ score and Race.

Identified subgroups can also guide the design of future studies. For example, if subgroups with differential treatment effects were identified by a GLMM tree in one randomized clinical trail (RCT), a next RCT may be designed so that it has adequate power to establish the differential subgroup effects with a confirmatory analysis. Or, after developing a new quiz, or improving the quiz analysed in Application Example 2, a validation study may be conducted so as to assess DIF between the subgroups identified with a Rasch tree in an earlier analysis. This would allow to assess whether the DIF has been mitigated in the new or improved quiz.

As mentioned in Application Example 1, resampling or k-fold cross validation techniques may be used to assess stability and accuracy of the GLMM tree. Philipp, Rusch, Hornik, and Strobl (2018) show how to use these techniques for assessing the stability of selected variables and cutpoints. For GLMM trees, cross-validation can additionally be used for evaluating predictive accuracy (e.g., de Rooij & Weeda, 2020), or for further optimizing predictive accuracy through tuning model-fitting parameters, like the minimum number of observations to be retained in terminal nodes or the maximum tree depth.

Application Example 2 showed that the Rasch tree method can employ an effect size measure to stop the tree from growing if effect-size differences are non-substantial. This also supports researchers in interpreting the tree's results with respect to whether DIF effects are negligible, medium, or large. For mixed-effects models, computation of effect sizes is much less straightforward and several different approaches have been proposed (e.g., Brandmaier, von Oertzen, Ghisletta, Lindenberger, & Hertzog, 2018; Brysbaert & Stevens, 2018; Judd, Westfall, & Kenny, 2017). Effect-size based stopping criteria for GLMM trees are therefore currently not available, but in principle, any effect size that can be computed for traditional GLMMs can also be computed for GLMM trees.

For GLMM trees, the use of cluster-level partitioning variables would commonly be desired in longitudinal analyses. As with traditional GLMMs, specification of the fixed and random effects requires careful consideration. For example, whether random slopes of time should be estimated, or whether random intercepts are sufficient to account for dependency between observations. Building on recent work by Fokkema and Zeileis (in press), we will further study and develop methods to take various dependence structures into account. For instance, in future implementations of `glmertree`, parameter stability tests that allow for detecting subgroup differences in the random-effects parameters (Wang, Merkle, Anguera, & Turner, 2021) may be implemented.

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. doi: 10.1111/j.1745-3984.1992.tb00368.x

Anthony, C. J., DiPerna, J. C., & Lei, P.-W. (2016). Maximizing measurement efficiency of behavior rating scales using item response theory: An example with the social skills improvement system—teacher rating scale. *Journal of School Psychology*, *55*, 57–69. doi: 10.1016/j.jsp.2015.12.005

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Lindenberger, U., & Hertzog, C. (2018). Precision, reliability, and effect size of slope variance in latent growth curve models: Implications for statistical power analysis. *Frontiers in Psychology*, *9*, 294. doi: 10.3389/fpsyg.2018.00294

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* New York: Chapman and Hall.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). doi: 10.5334/joc.10

Bureau of Labor Statistics, U.S. Department of Labor. (2019). National Longitudinal Survey of Youth 1979 cohort, 1979-2016 (rounds 1-27) [Computer software manual]. Columbus, OH.

Chung, G., Ansong, D., Brevard, K. C., & Chen, D.-G. (2021). Identifying treatment moderators of a trauma-informed parenting intervention with children in foster care: Using model-based recursive partitioning. *Child Abuse & Neglect*, *117*, 105065. doi: 10.1016/j.chiabu.2021.105065

De Ayala, R., & Santiago, S. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, *60*, 25–40. doi: 10.1016/j.jsp.2016.01.002

Debelak, R., Strobl, C., & Zeigenfuse, M. D. (2022). *An introduction to the Rasch model with examples in R.* Chapman & Hall/CRC. doi: 10.1201/9781315200620

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111–134.

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263.

Ding, C., & Zhao, Y. (2019). Using tree-based regression to examine factors related to math ability among 15-year old students. *Psychological Test and Assessment Modeling*, *61*(4), 453–470.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised.* Circle Pines, MN: American Guidance Service.

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2021). Generalized linear mixed-model (glmm) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, *31*(3), 329–341. doi: 10.1080/10503307.2020.1785037

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*, 2016–2034. doi: 10.3758/s13428-017-0971-x

Fokkema, M., & Zeileis, A. (in press). Subgroup detection in linear growth curve models with generalized linear mixed model (GLMM) trees. *Behavior Research Methods*.

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, *75*(2), 208–234. doi: 10.1177/0013164414536183

Frick, H., Strobl, C., & Zeiles, A. (2014). To split or to mix? Tree vs. Mixture models for detecting subgroups. In M. Gilli, G. González-Rodríguez, & A. Nieto-Reyes (Eds.), *Compstat 2014 – proceedings in computational statistics* (pp. 379–386). The International Statistical Institute/International Association for Statistical Computing.

Fritsch, M., Haupt, H., Lösel, F., & Stemmler, M. (2019). Regression trees and random forests as alternatives to classical regression modeling: Investigating the risk factors for corporal punishment. *Psychological Test and Assessment Modeling*, *61*(4), 389–417.

Henninger, M., Debelak, R., & Strobl, C. (2023). A new stopping criterion for Rasch trees based on the Mantel-Haenszel effect size measure for differential item functioning. *Educational and Psychological Measurement*, *83*(1), 181–212. doi: 10.1177/001316442210771

Holland, P. W., & Thayer, D. T. (1985). An alternate definition of the ETS delta scale of item difficulty. *ETS Research Report Series*, *1985*(2), i-10. doi: 10.1002/j.2330-8516.1985.tb00128.x

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., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*, 3905-3909.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625.

Komboz, B., Strobl, C., & Zeileis, A. (2017). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, *78*(1), 128–166. doi: 10.1177/0013164416664394

Kopf, J., Augustin, T., & Strobl, C. (2013). The potential of model-based recursive partitioning in the social sciences: Revisiting Ockam's Razor. In J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 75–95). New York: Routledge.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56.

Loh, W., & Shih, Y. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840.

Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology*, *35*(3), 299–314.

Philipp, M., Rusch, T., Hornik, K., & Strobl, C. (2018). Measuring the stability of results from supervised statistical learning. *Journal of Computational and Graphical Statistics*, *27*(4), 685–700.

Quinlan, J. R. (1993). *C4.5: Programms for machine learning.* San Francisco: Morgan Kaufmann Publishers Inc.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.).* Thousand Oaks, CA: Sage.

Stemmier, M., Heine, J.-H., & Wallner, S. (2019). Analyzing tree structures with configural frequency analysis and the r-package confreq. *Psychological Test and Assessment Modeling*, *61*(4), 419–433.

Strobl, C., Kopf, J., Kohler, L., von Oertzen, T., & Zeileis, A. (2021). Anchor point selection: Scale alignment based on an inequality criterion. *Applied Psychological Measurement*, *45*(3), 214–230. doi: 10.1177/0146621621990743

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316. doi: 10.1007/s11336-013-9388-3

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. doi: 10.5282/ubm/epub.10589

Strobl, C., Schneider, L., Kopf, J., & Zeileis, A. (2021). Using the raschtree function for detecting differential item functioning in the Rasch model [Computer software manual]. Vignette of R package `psychotree`.

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135–153. doi: 10.5282/ubm/epub.10588

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

van Wie, M. P., Li, X., & Wiedermann, W. (2019). Identification of confounded subgroups using linear model-based recursive partitioning. *Psychological Test and Assessment Modeling*, *61*(4), 365–387.

Wang, T., Merkle, E. C., Anguera, J. A., & Turner, B. M. (2021). Score-based tests for explaining heterogeneity in linear mixed models. *Behavior Research Methods*, *53*, 216–231. doi:

10.3758/s13428-020-01375-7

Wiedermann, W., Frick, U., & Merkle, E. C. (2021). Detecting heterogeneity of intervention effects in comparative judgments. *Prevention Science*, 1–11. doi: 10.1007/s11121-021-01212-z

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. doi: 10.1198/106186008X319331

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: an object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, *95*, 1–36. doi: 10.18637/jss.v095.i01

Appendix

Items of the culture scale for Application Example 2

1. Which painter created this painting? – Andy Warhol.

2. What do these four buildings have in common? – All four were designed by the same architects.

3. Roman numbers: What is the meaning of CLVI? – 156.

4. What was the German movie with the most viewers since 1990? – Der Schuh des Manitu.

5. In which TV series was the US president portrayed by an African American actor for a long time? – 24.

6. What is the name of the bestselling novel by Daniel Kehlmann? – Die Vermessung der Welt (Measuring The World).

7. Which city is the setting for the novel 'Buddenbrooks'? – Lübeck.

8. In which city is this building located? – Paris.

9. Which one of the following operas is not by Mozart? – Aida.