# One Model May Not Fit All: Subgroup Detection Using MOB

Marjolein Fokkema[1] and Mirka Henninger[2] & Carolin Strobl[2]

[1]Leiden University

[2]Universität Zürich

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 of intervention effects. 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: mixed-effects and item response theory models. 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 a dataset from a study of (Deming, 2009), who compared longitudinal performance of siblings who did, and siblings who did not participate in the Head Start program. Using GLMM trees, we identify subgroups of families in which children show comparatively larger or smaller gains in performance following participation in Head Start. In a second case study, we show how Rasch trees (Strobl, Kopf, & Zeileis, 2015a) 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 comparing groups, because undetected DIF can affect test fairness. We show how a recently developed stopping criterion (Henninger, Debelak, & Strobl, 2023a) can be used to guide subgroup detection based on DIF effect sizes.

Target journal: *Journal of School Psychology*. This Special Issue is on ”Conceptual and Methodological Advances for Understanding Contextual, Identity, and Cultural Effects in Intervention Research”. Submission deadline: October 31, 2023

## Introduction

Model-based recursive partitioning (MOB, Zeileis et al., 2008) is a semi-parametric approach for detecting differences in the parameters of a statistical model between groups of persons. This method is a generalization of the principle of recursive partitioning, that is also used in classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984). In classification and regression trees, the aim is to detect groups of persons, specified by (combinations of) covariates, that differ in a response variable. An example could be that children who are highly motivated and have good reading skills show higher average math exam grades. MOB generalizes this idea by identifying groups of persons, specified by (combinations of) covariates, that differ in a the parameters of a statistical model. An example could be that children who are highly motivated and have good reading skills show higher slopes in a regression model relating the time spent studying to math exam grades.

The framework of MOB is very flexible and can be applied in combination with a variety of statistical models, such as linear and generalized linear regression (Kopf, Augustin, & Strobl, 2013; Zeileis et al., 2008) or models for paired-comparison data (Strobl, Wickelmaier, & Zeileis, 2011). In this article, we will highlight two such combinations that we consider particularly relevant for research in school psychology: MOB for mixed-effects models (Fokkema et al., 2018) and for models from Rasch measurement and Item Response Theory (IRT, Henninger, Debelak, & Strobl, 2023b; Komboz, Strobl, & Zeileis, 2017; Strobl, Kopf, & Zeileis, 2015b). Mixed-effects models become relevant whenever data are collected in repeated measures or nested data structures, for example when children are tested at several time points (so that time points are nested in children) and/or when children from different classes from different schools (so that children are nested in classes, which are again nested in schools) participate in a study. In application example 1 we will illustrate how MOB can be used to detect subgroup-specific intervention effects while taking into account the nested data structure.

While for this first application example we assume that the administered test has already been thoroughly validated, for application example 2 we go back in the research process to the point where a new test is administered to a validation sample. Before using the test in our research, we need to make sure that test results are comparable, for example between children of different genders. If certain items pose an advantage or disadvantage to either group, they 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;

should we call a psychological test a test/an instrument/??? to distinguish it from the verb and noun for statistical test? (Mirka: I like the

Maller, 1997), groups of persons are compared with respect to the item parameters for their specific group. This can be done in a way that allows to detect DIF on top of any true group differences in ability. While many standard approaches for testing DIF are limited to comparing typically two pre-specified groups at a time, we will show in application example 2 that MOB more flexibly allows to detect groups of persons with different item parameters in a data-driven way.

In the following section, we will 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). An alternative framework for detecting 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 covariates. A comparison of both approches is given by Frick, Strobl, and Zeiles (2014).

## The Algorithm Behind MOB

What you see displayed in Figures 2, **??**, and **??**, which will be explained in detail in the application example sections, is the result of the MOB algorithm for a linear mixed-effects model (Figure 2) and for a binary Rasch model (Figures **??** and **??**). These figures share the tree-structure in the upper part, but differ in their lower part. The starting point of the tree structure is at the very top of each tree. The very top of each tree is called the root node, so what you see are actually "upside down" trees with the root at the top and the branches at the bottom. The root node at the very top contains the entire sample of persons in the data set. From there, the persons are divided into subgroups.

adjust if raschtree example changes (Mirka: maybe it would be helpful to have a simplified figure of a GLMERtree and a Raschtree early in the manuscript already, so we can refer to it here?)

The subgroups are defined by the covariates that are used for splitting and together with cutpoints in those covariates. For example, in Figure 2 for the metric covariates `AFTQ` and `Income` a numeric cutpoint has been selected by the MOB algorithm to separate the groups. The algorithm has identified this cutpoint as being the location of the strongest parameter difference between the two resulting groups. The categorical covariate `Race`, which was coded in three categories in this example, has been divided into two groups of categories: Black and Hispanic vs. White. For a binary covariate, on the other hand, there is only one possible cutpoint, namely between the two categories. While it is also possible to create more than two groups in each split (Kim & Loh, 2001; Quinlan, 1993), binary splitting algorithms are typically preferred because they lead to more concise trees.

like ... – add name if is there one in the other example?

The parametric model to which the MOB algorithm is applied is visualized in the end nodes at the bottom of each tree. While in Figure 2 each end node contains a linear mixed-effecs model with a group-specific effect for the Head Start intervention, in Figures ?? and ?? each end node contains the group-specific item parameters of a Rasch model for the respective test items. The fact that the Figures display more than one end node already means that one joint mixed-effects or Rasch model was not appropriate to describe the pattern in the data. Figure 2 also shows that, out of the covariates that were presented to the algorithm (AFTQ, Race, Income,Mom_edu_yrs, and Mom_height), only AFTQ, Race, and Income were actually selected for splitting. Variables are selected based on a statistical test for parameter differences, also termed a test for structural change. Structural change is present if, for example, the model parameters systematically differ for children from lower vs. higher income families. At each node, the algorithm will select the covariate showing the strongest structural change as the next splitting variable. Within this variables, the optimal cutpoint is chosen in a separate step.

The way the variable selection and cutpoint selection steps are separated in modern classification and regression tree and MOB algorithms, including the ones used here, distinguish them from the traditional classification and regression tree algorithms of Breiman et al. (1984) and Quinlan (1993). The traditional algorithms performed variable and cutpoint selection in one step. However, this leads to an undesirable behavior called variable selection bias. It means that the traditional algorithms prefer variables offering more 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). Modern algorithms for classification and regression trees that have solved this problem are the unbiased approaches QUEST (Loh & Shih, 1997), which is available in SPSS, and ctree (Hothorn, Hornik, & Zeileis, 2006), available in R in the party and partykit packages (Hothorn & Zeileis, 2015). The latter forms the basis for all MOB approaches presented in this paper.

Once there are no more covariates that show a significant structural change in any node, the MOB algorithm stops splitting. In this way, the MOB algorithm selects only those variables that are relevant for distinguishing the groups, i.e., it performs automatic variable selection. Moreover, the trees will not grow as large as possible, but will stop when no more significant structural change is detected. This is the second difference between the modern classification and regression tree and MOB algorithms used here, compared to traditional classification and regression tree algorithms like those of Breiman et al. (1984) and Quinlan (1993): While the traditional algorithms grew very large trees and then cut them back by means of so called pruning, the framework employed here is based on statistical significance tests and partly also on effect size measures (see application example 2) to stop the trees already in the growing phase when no more significant change can be detected. Other stopping criteria are based on the number of persons in the end nodes. These crieria ensure that the sample sizes in the end nodes are large enough to estimate the statistical model in each end node.

we should discuss this in the application examples; there are some defaults but for models with many parameters it might make sense to increase them

In summary, the rationale of a MOB algorithm consists of the following steps:

why are the spaces so big? I didn't change the enumerate command

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

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

3. If there is significant structural change, groups are split along the covariate with the strongest change and using the optimal cutpoint.

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

For more details on the statistical theory behind unbiased classification and regression trees and MOB, see Hothorn et al. (2006); Strobl et al. (2015a, 2009).

An important 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 way, but is learned from the data in an exploratory way. This is a key feature of the MOB approach that makes it very flexible and distinguishes it 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 hypothesis a priori and test them in a confirmatory way, in early stages of research exploratory methods are an important addition to the statistical toolbox for researchers.

**Previous Papers on IRT and Rasch Models in Journal of School Psychology**
Rasch and IRT models are often mentioned in this journal's papers as being used for scoring school assessments (e.g., Virginia Standards of Learning Assessments; Oregon Assessment of Knowledge and Skills; Peabody individual achievement test, Luther (1992)).
A general introduction to IRT in the journal is provided by Anthony et al. (2016). included There are also papers describing extensions to mixure IRT models De Ayala and Santiago (2017) included and many-facet Rasch measurement (MFRM) approaches that allow controlling for rater effects Styck, Anthony, Flavin, Riddle, and LaBelle (2021).
Maller (1997) assessed DIF by comparing Rasch model item difficulties of WISC-III subtests between 110 severely and profoundly deaf children, 110 matched nonreferred hearing children, and the WISC-III standardization sample (N = 2,200). included

The general framework of MOB introduced above will now be applied to two types of statistical models that are particularly relevant in school psychology research: Mixed-effects models for repeated measures or nested data structures and measurement models for validating psychological and educational tests.

## Using MOB for Subgroup Detection in Mixed-Effects Models

In mixed-effects models contain two types of effects: fixed effects and random effects. Fixed effects are typically used, for example, to describe the effect of an intervention on the average student performance. Random effects are used to describe the variation in the intercept and slope for, e.g., students in different classes.

When mixed-effects models are combined with MOB, ...

Marjolein please edit and finish

## Using MOB for Detecting DIF in Measurement Models

Before being able to use a psychological or educational test for comparing different groups of persons in a fair way, different assumptions of the measurement model have to be checked. In the framework of IRT and Rasch measurement, test items are typically investigated with respect item misfit, multidimensionality and other violations of the measurement model (cf., e.g., Debelak et al., 2022, for an introduction). Another assumption that is particularly crucial 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 display DIF.

MOB can be used to detect DIF by means of searching for covariates that display structural change in the item parameters of the measurement model. For the Rasch measurement model, the R function for conducting the MOB analysis is the `raschtree` function from R package `psychotree` (Strobl et al., 2015a) The usage of this function will be illustrated in application example 2. We will see that, just like for the mixed-effects model 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 variables are made available to the MOB algorithm for selecting relevant splitting variables and cutpoints.

should we cite the package too or only the paper?

use same names for methods everywhere

If one joint Rasch model holds for the entire sample, i.e., if there is no DIF, a Rasch tree should show no splits. However, in certain settings in educational research, such as large scale assessements, very large sample sizes are available for testing for DIF. Large sample sizes are good for detecting even small effects or model violations with a 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 with large samples. 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. As we will illustrate in the first part of application example 2, for a large sample of university students who have taken a quiz to test their general knowledge, this can lead to very large trees, that are hard to interpret and contain splits that would not be considered relevant by measurement experts. Therefore, an extension of Rasch trees has been suggested by Henninger et al. (2023a) 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. (2023a) 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.

As we will show in application example 2, for large sample sizes this can be very helpful because it results in shorter trees that are easier to interpret and contain only splits corresponding to DIF effect sizes considered relevant in practice. Together with a purification step (see Henninger et al., 2023a, for details), the Mantel-Haenszel classification can also be used for highlighting those items that show DIF with respect to certain groups of persons graphically. This can help generate hypothesis about the sources of DIF, as we will ilustrate in application example 2, and can also aid the decision how to proceed with the DIF items.

following text from Caro can be used here or later after example 2 or in discussion

For example, items that show DIF between different language groups can often be improved by means of making sure that in all translations the meaning is as similar as possible, that the words employed in the translations for the different languages are equally frequently used, etc. In other situations, sources of DIF might be harder to eliminate, so that often DIF items are 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 would be the case in instruments intended to measure math aptitude containing 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 they contain seldomly used words. These items will 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, e.g., by using more frequently used words, or whether the test should be considered two-dimensional rather than one-dimensional (see also Ackerman, 1992; Strobl, Kopf, Kohler, von Oertzen, & Zeileis, 2021, for a discussion of the connection between DIF and multidimensionality).

We will now illustrate how to use the R packages `glmertree` and `psychotree` to conduct the MOB analyses.

## Application Example 1: Subgroup Detection in Mixed-Effects Models

### Dataset

To illustrate the use of GLMM trees, we analyze a dataset from Deming (2009), who evaluated long-term benefits of participation in Head Start. Head Start is a federally funded nationwide pre-school program for children from low-income families. Deming (2009) compared performance of siblings who differed in their participation in the program using data from the National Longitudinal Survey of Youth (NLSY; REF).

what nation?

The sample consists of 273 families with at least two children, where at least one child participated in Head Start and at least one child did not participate in Head Start or any other preschool program. Data from children in those families who participated in another preschool program were excluded. As such, siblings who did not participate in Head Start serve as a natural control to assess the effects of participating in Head Start. As the outcome variable, we take the Peabody Picture Vocabulary Test (PPVT; REF) and model PPVT trajectories over time. As the timing metric, we took the square root of the child's age in years, which yielded a pattern of approximately linear increases over time.

this would scare me a little as a reader: how would I know what to do/that I should do this? (Mirka: Maybe we can add that the sqrt or the log is a typical way to model trajectories that first increase rapidly, but then later increase more slowly?)

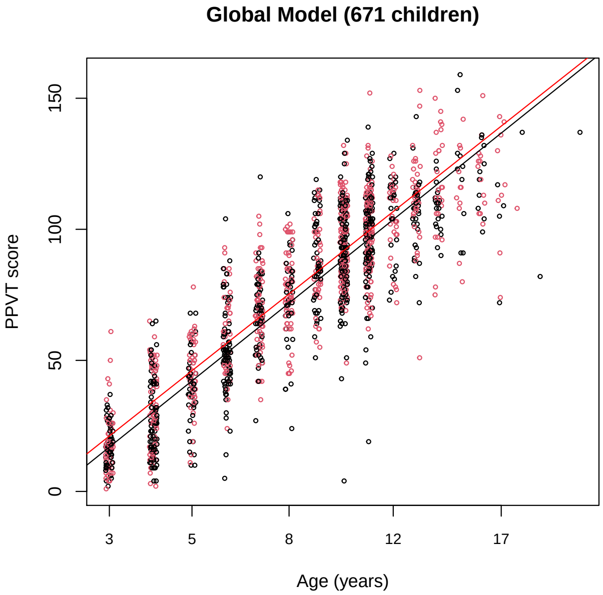
Our dataset contains five family characteristics: The mother's score on the Armed Forced Qualification Test (AFTQ), adjusted for age; the families' income (averaged over the years for which data was available); race (Black, Hispanic or White); mother's years of completed education; mother's height. Note 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 comprises data from families and children for whom complete data was available.

We load the data and inspect the first rows:

```
HS_dat <- readRDS("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
1   18 2.000000
2   48 2.645751
3   69 2.645751
```

**Figure 1**



We inspect the complete dataset by plotting PPVT scores against age, separated by program participation: None (black) versus Head Start (red). To show the effect of age and Head Start participation, we fitted a mixed-effects model comprising their main and interaction effects. To account for the correlation between repeated assessments on the same child, we specified a random intercept. The results are presented in Figure 1, which shows that children participating in Head Start show slightly higher performance than their non-participating siblings and that this difference is slightly diminished but persists over time. This result agrees with the findings of Deming (2009).

code for figure not displayed on purpose? why only those particular x axis ticks?

## Linear mixed effects model tree

I (Mirka) am confused: the nested structure in this example is timepoints (age) within children? Therefore, ChildID is the cluster-variable. But then, there should be residual errors between the two children from the same family? In the text it says that this is accounted for by the random intercept, but I don't think this is the case (but maybe I am wrong?). Is it maybe accounted for by the cluster argument that is also said to ChildID? To me, the model specification is not yet clear from the text. Plus: If I am right, and the cluster is the child and the timepoints are on Level-1, I think all the covariates are on Level-2 (child level), and then maybe we should comment on the Type-I error problem?

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|ChildID) | AFTQ + Race +
                    Income + Mom_edu_yrs + Mom_height,
                  data = HS_dat, cluster = ChildID, 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|ChildID`) specified the random effects. 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.

(Mirka) Maybe add information that this is for each child, average DV across timepoints

I (Mirka) am not sure what exactly the cluster argument does. Maybe Marjolein you have more information from your work with Achim. But my knowledge/intuition is that it has something to do with clustered covariances, so maybe it accounts for the sibling structure, but I don't think it accounts for the level of the predictor variables.

With the second argument, we specified the dataset which contain the variables. Because we are dealing with repeated measurements on the same children, we additionally specified that the parameter stability tests should be performed on the child level using the `cluster` argument. Using the default observation-level parameter stability tests may artificially inflate power. Finally, because we want to retain large enough subgroups, we specified that the minimum number of observations in a terminal node should be 250.

Mirka: I'd rather say it has an inflated type-I error rate.

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, not the number of individual observations:

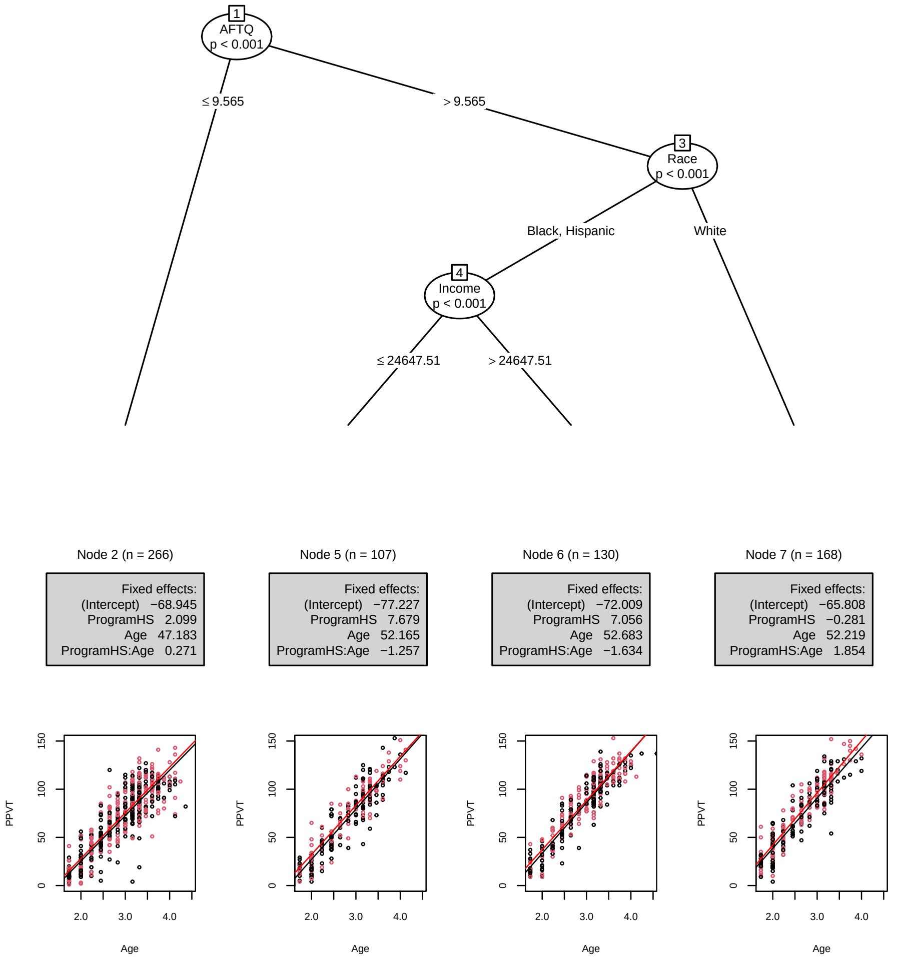
```
plot(HS_tree, type = "simple", nodesize_level = 2)
```

The resulting tree is presented in Figure 2. 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 income.

To aid interpretation of the coefficients in Figure 2, Table 1 provides predicted PPVT scores for each of the groups and programs at ages 6 and 18. All nodes show a modest benefit

x axis for age starts at 2?

## Figure 2



## Table 1

*Predicted PPVT scores at different ages.*

| Node | Program | PPVT at age 6 | PPVT at age 18 |
|---|---|---|---|
| 2 | None | 46.63 | 131.23 |
| 2 | HS | 49.39 | 134.48 |
| 5 | None | 50.55 | 144.09 |
| 5 | HS | 55.15 | 146.43 |
| 6 | None | 57.04 | 151.51 |
| 6 | HS | 60.09 | 151.63 |
| 7 | None | 62.10 | 155.74 |
| 7 | HS | 66.36 | 163.32 |

**Table 2**

*Cross-validated performance of LMMs and LMM trees.*

| method | MSE | SD | number.of.splits | R2 |
|---|---|---|---|---|
| LMM tree | 221.022 | 55.727 | 2.7 | 0.813 |
| LMM | 258.233 | 71.451 | NA | 0.781 |

of Head Start participation at age 6, about 3-4 points on the PPVT. This benefit remains the same over time for the group with lower mother's AFTQ scores. The benefit increases over time for White children with higher mother's AFTQ scores. Strikingly,the benefit decreases over time for Black and Hispanic children with higher mother's AFTQ scores. These results correspond to the conclusions of Deming (2009).

possibly less extreme word like: Interestingly,

Generally the effects are not large but the example is very clear and I like that it is an intervention effect. I suggest to state early before the figure that the effects are not large in this example but we are interested in the group differences; also try to make figure bigger and wider to make slopes more visible; (Mirka): I think the main effect of the intervention and the interaction effect is not too small, but it is hard to see because the age effect is so pronounced in this example. Would it be an option to only use the intervention effect in the end nodes?

To evaluate whether the detected subgroups indeed contribute to better predictions, we used cross validation. We randomly separated the 258 families in the dataset into ten equally-sized folds. We took nine of the ten folds as a training dataset on which we fitted two models: an LMM tree like the one shown above (i.e., an LMM with subgroups ) and an LMM without subgroups (comprising main and interaction effects of age and Head Start participation, and a random intercept with respect to child). We evaluated performance on the remaining folds by computing and evaluating accuracy of the predictions. We repeated this procedure ten times, so that all folds were used as a test set once. The results are presented in Table 2, which shows that LMM trees generalize well: They provide better predictive accuracy compared to LMMs, while implementing only few splits.

necessary? confused me

explain MSE and $R^2$

I am not sure this dataset is the best GLMM tree illustration, because visualized differences between subgroups are very small and perhaps the fixed-effect part is already too complex. Alternative: Do not focus on effect of Head Start, but use as partitioning variable. Then can also use data from more families (HS or none not needed), partition using child-level characteristics, and model specification becomes simpler. – response Caro: I like it, see comment above: try to prime that effects are not large – response Mirka: I think that's okay. Only if you had data at hand that is not longitudinal (rather students in classes/schools), with covariates on Level-1, I would opt for that because it doesn't have the Level-2 splitting issue.

## Application Example 2: Subgroup Detection in Rasch Models

- Example Rasch tree: Data from SPISA (data from Trepte & Verbeet)
- take the subsample of students (only students), and a random subsample of 5000 respondents (seed 04102023)
- items could probably appear in Appendix??
  - Which painter created this painting? – Andy Warhol.
  - What do these four buildings have in common? – All four were designed by the same architects.
  - Roman numbers: What is the meaning of CLVI? – 156.
  - What was the German movie with the most viewers since 1990? – Der Schuh des Manitu.
  - In which TV series was the US president portrayed by an African American actor for a long time? – 24.
  - What is the name of the bestselling novel by Daniel Kehlmann? – Die Vermessung der Welt (Measuring The World).
  - Which city is the setting for the novel 'Buddenbrooks'? – Lübeck.
  - In which city is this building located? – Paris.
  - Which one of the following operas is not by Mozart? – Aida.

at Caro: Ich habe die Daten (glaube ich) mal von dir bekommen (`spisa_ges.RData`) und die Item-Formulierungen habe ich aktuell aus dem psychotree Paket. Stimmt das so? Ich glaube, der Originaltest hatte mehrere parallele Skalen, zumindest sind in dem PDF/BUCH von Trepte sehr viel mehr, und sehr ähnliche Items enthalten. Das könnten wir vielleicht nochmal nachgucken.

We load the data and inspect the first rows. there are several covariates: age (continuous), gender (male, female, missing), area (Language & Culture, Law & Economics, Medicine & Health, Engineering, Sciences, Pharmacy, Geography, Agriculture & Nutrition, Sports) and 9 items from the culture scale (true/false 0/1).

```
head(covar_data)
```

```
       Age Gender                Area
445472  28   Male  Language & Culture
167408  26 Female         Engineering
422161  25   Male  Language & Culture
```

```
350734  27   Male         Engineering
570206  23   Male \nMedicine & Health
265570  21 Female                Arts
```

```
 head(culture_scale)
```

```
     i27 i28 i29 i30 i31 i32 i33 i34 i35
[1,]   1   1   0   0   1   1   1   1   1
[2,]   1   0   1   0   0   0   1   1   1
[3,]   1   1   1   0   1   1   1   1   1
[4,]   0   1   0   1   0   1   1   0   1
[5,]   1   1   0   1   1   1   0   1   0
[6,]   1   1   0   1   1   0   1   1   1
```

We create a dataset `dat_SPISA` that contains the data of the covariate and the item responses from the culture scale. By assigning the whole scale to the dataset using `$`, we can later access item responses from all items using `$culture`.

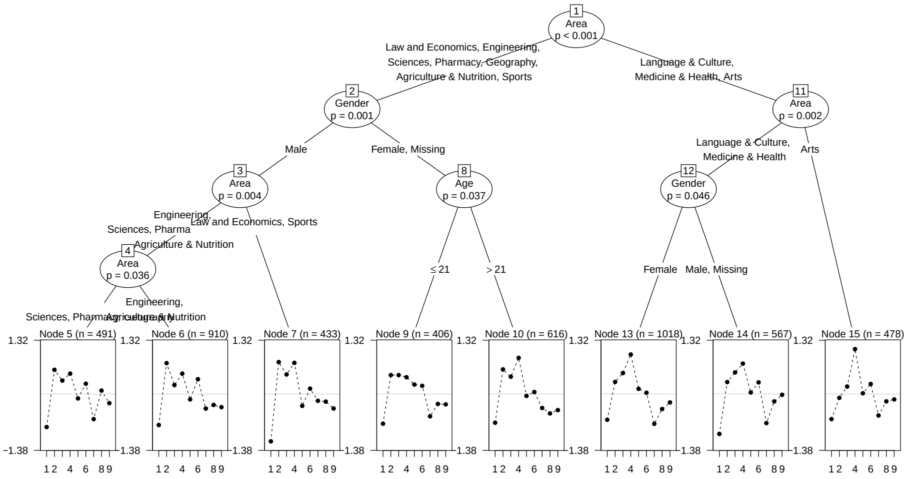
```
 dat_SPISA <- covar_data
 dat_SPISA$culture <- culture_scale
```

The Rasch tree can be fitted with the function `raschtree` from the R package `psychotree`. We use the typical formula syntax with the item responses of the culture scale on the left hand side of ∼, and the covariates on the right hand side.

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

We can plot the Raschtree using the plot function. But it is very big, lots of splits, we don't know whether the splits are due to substantial differences in item difficulty parameters, because the item difficulty profiles in the end nodes are hard to interpret/compare.

```
 plot(Raschtree_culture)
```



Mantel-Haenszel trees use the Mantel-Haenszel effect size measure for DIF. it has three categories (A: negligible, B: moderate, C: large). If none of the items has DIF in category B or C (default, but can also be changed so that DIF has to be C), the tree is stopped from growing.

The software must be installed from github, but will (probably) be implemented in the psychotree package some time in the future

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

Syntax is very similar to the classical Rasch tree, but an additional argument `stopfun`, where the mantelhaneszel stopping function can be selected (but also own stopping functions can be provided; see Henninger et al., 2023). The user also has to indicate what kind of purification strategy should be used (none, 2step, iterative). Iterative is recommended.

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

For technical reasons, the information about the effect size and ETS classification are not saved in the tree object itself but have to be added afterwards using the `add_mantelhaenszel` function. The information can then be accessed from the object using `$info$mantelhaenszel`

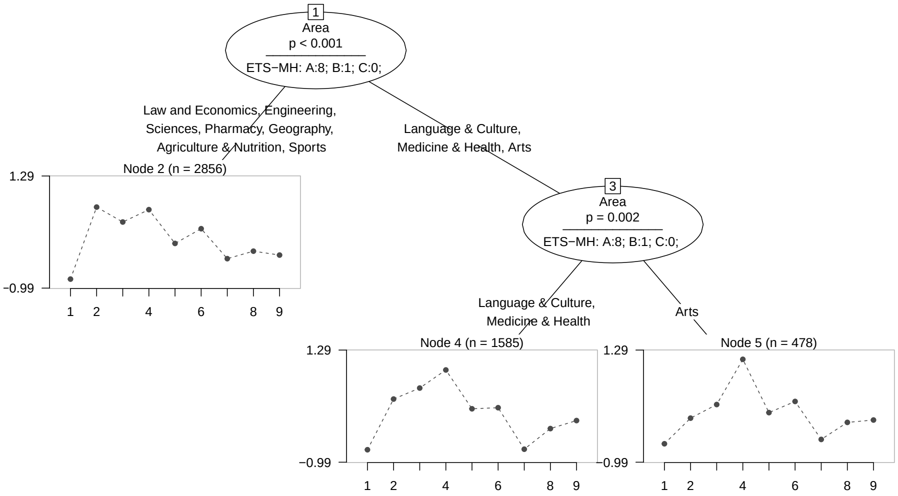
```
Raschtree_MH_culture <- add_mantelhaenszel(Raschtree_MH_culture,
                                           purification = "iterative")
```

```
Raschtree_MH_culture$info$mantelhaenszel
```

not clear what this information is necessary/useful for? since it is not displayed, leave this out?

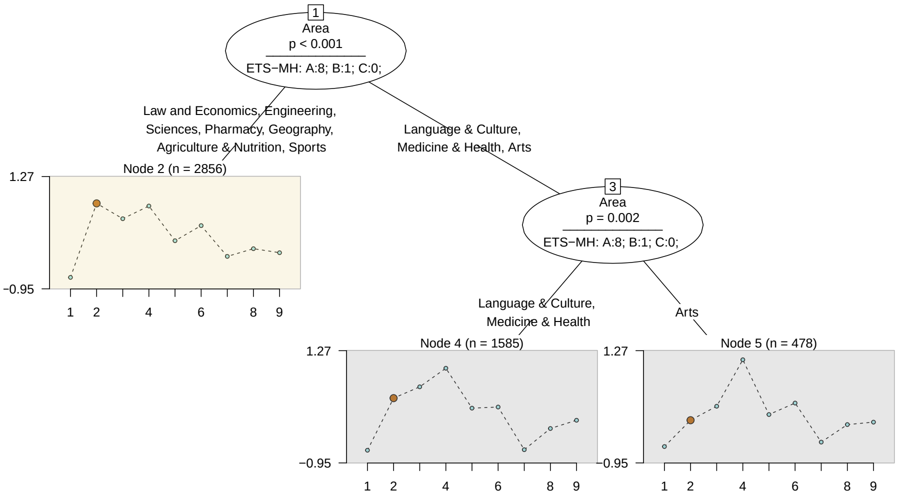
If we now plot the Mantel-Haenszel Rasch tree, we see that the tree is much more concise with a lower number of splits. In addition, we see the number of items classified as A, B, or C in each node. For instance, we see that one item is classified as B in Node 1, and no item as C.

```
plot(Raschtree_MH_culture)
```



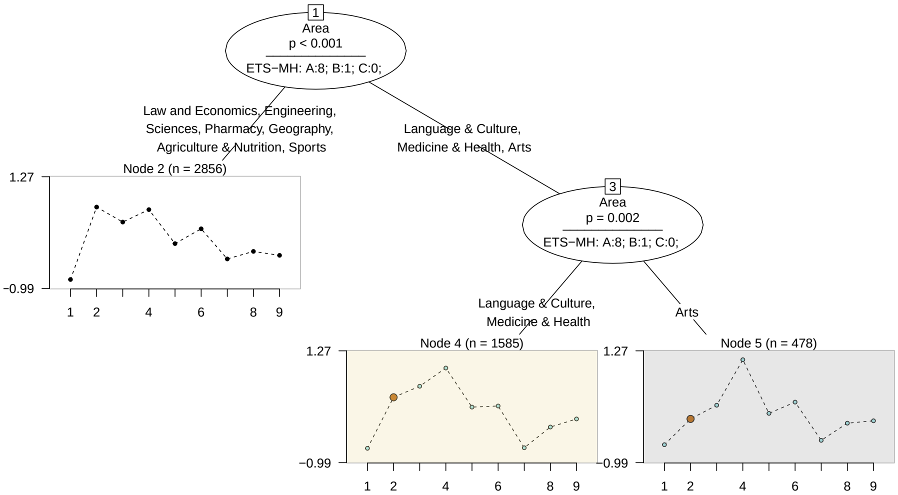
We can also color the items in the end node profiles according to a split in the tree. Here, we see that for Node 1 (split on the covariate Area), item 2 shows DIF in category B. Item 2 seems to be easier (y-axis shows item easiness) in certain areas compared to Language, Culture, Medicine, Health, Arts.

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



We can also color by Node 3 (also split on covariate Area). Also here, Item 2 shows DIF. Similarly, item 2 seems to be more difficult (less easy) for respondents who study Arts.

```
plot(Raschtree_MH_culture, color_by_node = 3)
```



This is interesting, because it is the culture scale. Why should people who study Arts have more problems to solve the item? When we look at the item, it probably makes sense: the items asks what four buildings have in common (Allianz-Arena München, Tate Modern NY, Olympia-Stadion Peking, Elbphilharmonie Hamburg). Two of these buildings are sports stadiums. So probably respondents who are interested in sports, and know these buildings well have an advantage on the item (independent of the level of knowledge about culture).

Maybe a small note on anchoring? Item difficulty profiles are anchoring on the items that do not show DIF. This facilitates comparisons across end nodes (but maybe we do not want to say that??)

## 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 way. This makes it more flexible for detecting differences that were not hypothesized by the researchers. For example in DIF analysis, it is often the case that any obvious sources of DIF have already been avoided by the content experents in the item creation phase. Any remaining DIF is unexpected. Therefore, any DIF detection approach that relies on the researchers correctly specifying the exact groups of persons that exhibit DIF can miss DIF if it is associated with other (combinations of) covariates or other cutpoints than the ones investigated. In this sense, a more exploratory approach like a Rasch tree has a higher statistical power to detect DIF in previously unknown groups (Strobl et al., 2015b). The same holds for parameter differences in mixed-effects models, where the substantial hypothesis are typically about the fixed effects of the interventions, not necessarily about all possible subgroup-specific ....

Marjolein add if this makes any sense here

Write summary of findings.

For our two application examples, we could show that MOB can detect ..., performs automated variable selection

or did you, Marjolein, mean repeat substantive findings? not very exciting for SPISA

Mention shortcomings.

not possible to detect certain patterns: XOR problem, known for all trees

stability of trees, small changes in data can lead to quite different looking tree. The tree structure should always be interpreted jointly for all variables: rather than considering the first splitting variable to be the most important one, it is actually the combination of all splitting variables that creates the groups.. The **stabletree** R package also provides descriptive statistics and graphics that help judge the stability of tree and MOB results (Philipp, Rusch, Hornik, & Strobl, 2018; **?**).

in interpretation of trees in examples: do this, point out that it is interactions – Caro also add in intro?

Write about future work.

sample size dependence, similar ideas like MH trees also possible for other MOB?

The Mantel-Haenszel effect size measure for DIF will be integrated in the `psychotools` package.

## 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.

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.

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

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

Debelak, R., Strobl, C., & Zeigenfuse, M. D. (2022). *An introduction to the Rasch model with examples in R.* Chapman & Hall/CRC. Retrieved from `https://www.taylorfrancis.com/books/mono/10.1201/9781315200620/introduction-rasch-model-examples-carolin-strobl-matthew-zeigenfuse-rudolf-debelak` 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.

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.

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.

Henninger, M., Debelak, R., & Strobl, C. (2023a). 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.

Henninger, M., Debelak, R., & Strobl, C. (2023b). 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. Retrieved from `https://journals.sagepub.com/doi/10.1177/00131644221077135` doi: 10.1177/00131644221077135

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: https://doi.org/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. Retrieved from `https://jmlr.org/papers/v16/hothorn15a.html`

Kim, H., & Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*(454), 589–604.

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.

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

Luther, J. B. (1992). Review of the Peabody Individual Achievement Test-Revised. *Journal of School Psychology*, *30*(1), 31–39.

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. doi: 10.1080/10618600.2018.1473779

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

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. (2015a). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316.

Strobl, C., Kopf, J., & Zeileis, A. (2015b). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316. doi: 10.1007/s11336-013-9388-3

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., 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

Styck, K. M., Anthony, C. J., Flavin, A., Riddle, D., & LaBelle, B. (2021). Are ratings in the eye of the beholder? a non-technical primer on many facet Rasch measurement to evaluate rater effects on teacher behavior rating scales. *Journal of School Psychology*, *86*, 198–221.

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

Add DOIs.

## Appendix

### Appendix A: SDQ-II items administered in the ECLS

Fix formatting.

How true is each of these about you? (1=”not at all true”; 2=”a little bit true”; 3=”mostly true” or 4=”very true”)

- C7MTHBST (math): Math is one of my best subjects.
- C7ANGRY (internalizing): I feel angry when I have trouble learning.
- C7LIKRD (reading): I like reading.
- C7WRYTST (internalizing): I worry about taking tests.
- C7MTHGD (math): I get good grades in math.
- C7LONLY (internalizing): I often feel lonely.
- C7ENGBST (reading): English is one of my best subjects.
- C7SAD (internalizing): I feel sad a lot of the time.
- C7LIKMTH (math): I like math.
- C7WRYWEL (internalizing): I worry about doing well in school.
- C7ENJRD (reading): I enjoy doing work in reading.
- C7WRYFIN (internalizing): I worry about finishing my work.
- C7ENJMTH (math): I enjoy doing work in math.
- C7WRYHNG (internalizing): I worry about having someone to hang out with at school.
- C7GRDENG (reading): I get good grades in English.
- C7ASHAME (internalizing): I feel ashamed when I make mistakes at school.