Austrian Journal of Statistics
2025, Volume 54, 9–26.
http://www.ajs.or.at/
doi:10.17713/ajs.v54i3.2055

Examining Exams Using Rasch Models and Assessment of Measurement Invariance

Achim Zeileis
Universität Innsbruck

Abstract

Many statisticians regularly teach large lecture courses on statistics, probability, or mathematics for students from other fields such as business and economics, social sciences, psychology, etc. The corresponding exams often use a multiple-choice or single-choice format and are typically evaluated and graded automatically, either by scanning printed exams or via online learning management systems. Although further examinations of these exams would be of interest, these are frequently not carried out. For example a measurement scale for the difficulty of the questions (or items) and the ability of the students (or subjects) could be established using psychometric item response theory (IRT) models. Moreover, based on such a model it could be assessed whether the exam is really fair for all participants or whether certain items are easier (or more difficult) for certain subgroups of students.

Here, several recent methods for assessing *measurement invariance* and for detecting *differential item functioning* in the Rasch model are discussed and applied to results from a first-year mathematics exam with single-choice items. Several categorical, ordered, and numeric covariates like gender, prior experience, and prior mathematics knowledge are available to form potential subgroups with differential item functioning. Specifically, all analyses are demonstrated with a hands-on R tutorial using the *psycho** family of R packages (**psychotools**, **psychotree**, **psychomix**) which provide a unified approach to estimating, visualizing, testing, mixing, and partitioning a range of psychometric models.

The paper is dedicated to the memory of Fritz Leisch (1968–2024) and his contributions to various aspects of this work are highlighted.

Keywords: multiple choice, item response theory, differential item functioning, psychometrics, R.

1. Introduction

1.1. Large-scale exams

Statisticians often teach large lecture courses with introductions to statistics, probability, or mathematics in support of other curricula such as business and economics, social sciences, psychology, etc. Due to the large number of students and possibly also of lecturers who teach lectures and/or tutorials in parallel, it is often necessary to rely on exams and other assessments based on large pools of so-called closed (as opposed to open-ended) items, i.e., exercises which can be evaluated and graded automatically.

The most widely-used item type for such assessments are multiple-choice (also known as multiple-answer) or single-choice exercises. But due to the widespread adoption of learning management systems such as Moodle (Dougiamas *et al.* 2024), Canvas (Instructure, Inc. 2024), or Blackboard (Anthology Inc. 2024), especially since the Covid-19 pandemic, other item types are also being used increasingly. Often the evaluation is binary (correct vs. incorrect) but scores with partial credits for partially correct items are also frequently used.

1.2. Examining exams

Traditionally, mostly simple summary statistics have been used to assess the results from such large-scale exams, e.g., the proportion of students who correctly solved the different items and the number of items solved per student. However, recently there has also been increasing interest in so-called *learning analytics* which connects the results from different exams or assessments with covariates such as the field and duration of the study, prior knowledge from previous courses, etc., in order to better understand and shape the learning environments for the students (see Wikipedia 2024, for an overview and further references).

However, in Austria, to the best of our knowledge, it is still not common to apply standardized and/or automated psychometric assessments to exam results. A notable exception is the multiple-choice monitor at WU Wirtschaftsuniversität Wien introduced by Nettekoven and Ledermüller (2012, 2014). In addition to various exploratory techniques they also employ probabilistic statistical models from psychometrics to gain more insights into exam results. More specifically, they use models from item response theory (IRT, Fischer and Molenaar 1995; Van der Linden 2016), including the Rasch model (Rasch 1960) which is also employed in the analysis of international educational attainment studies such as PISA (Programme for International Student Assessment, `https://www.oecd.org/pisa/`).

1.3. Measurement invariance in IRT

Based on an exam's item responses, IRT models can estimate various quantities of interest, most importantly the *ability* of the individual students and the *difficulty* of the different items (or exercises). A fundamental assumption is that the models' parameters are invariant across all observations, which is also known as *measurement invariance* (see Horn and McArdle 1992, for an early overview in psychometrics). Otherwise observed differences in the items solved cannot be reliably attributed to the latent variable that the model purports to measure.

Typical sources for violation of measurement invariance in IRT models are *multidimensionality* (i.e., more than one latent variable instead of a single ability) or *differential item functioning* (see Debelak and Strobl 2024). The latter refers to the situation where the same item can be relatively easier or more difficult (compared to the remaining items) for different students, despite having the same latent ability.

1.4. Our contribution

In Section 2, we introduce a data set from one of our own mathematics courses, containing binary responses (correct vs. incorrect) from 13 items in an end-term exam of an introductory mathematics course for economics and business students. In Section 3, the Rasch model is briefly introduced, fitted to the data, and interpreted regarding the items' difficulties and students' abilities. Subsequently, in Section 4, various methods for capturing violations of measurement invariance are applied: (1) Classical two-sample comparisons of two exogenously given groups along with modern methods for anchoring the item difficulty estimates. (2) Rasch trees based on generalized measurement invariance tests for data-driven detection of subgroups affected by DIF. (3) Rasch finite mixture models as an alternative way of data-driven characterization of DIF clusters. Section 5 wraps up the paper with a discussion and the epilogue Section 6 concludes the paper by highlighting Fritz Leisch's influence on different aspects of this work.

In all sections, emphasis is given to the hands-on application of the methods in R (R Core Team 2024) – notably using the packages **psychotools** (Zeileis, Strobl, Wickelmaier, Komboz, Kopf, Schneider, and Debelak 2024a), **psychotree** (Zeileis, Strobl, Wickelmaier, Komboz, Kopf, Schneider, Dreifuss, and Debelak 2024b), and **psychomix** (Frick, Leisch, Strobl, Wickelmaier, and Zeileis 2024) – along with the practical insights about the analyzed exam.

2. Data: Mathematics 101 at Universität Innsbruck

The data considered for examination in the following sections come from the end-term exam in our "Mathematics 101" course for business and economics students at Universität Innsbruck. This is a course in the first semester of the bachelor program and it is attended by about 600–1,000 (winter) or 200–300 (summer) students per semester.

Due to the large number of students in the course, there are frequent online tests carried out in the university's learning management system OpenOlat (frentix GmbH 2024) as part of the tutorial groups, along with two written exams. All assessments are conducted with support from R package **exams** (Grün and Zeileis 2009; Zeileis, Umlauf, and Leisch 2014) which allows to automatically generate a large variety of similar exercises and render these into many different output formats.

In the following, the individual results from an end-term exam are analyzed for 729 students (out of 941 that had registered at the beginning of the semester). The exam consisted of 13 single-choice items with five answer alternatives, covering the basics of analysis, linear algebra, and financial mathematics. Due to the high number of participants, the exam was conducted with two groups, back to back, using partially different item pools (on the same topics). All students had individual versions of their items generated via R/**exams**. Correctly solved items yielded 100% of points associated with an exercise. Items without correct solution can either be unanswered (0%) or have an incorrect answer (−25%). In the following, the item responses are treated as binary (correct vs. not correct).

The data are available in the R package **psychotools** as `MathExam14W` where `solved` is the main variable of interest. This is an object of class '`itemresp`' which is internally essentially a 729×13 matrix with binary 0/1 coding plus some metainformation. To generate such '`itemresp`' objects from a numeric matrix the eponymous function `itemresp()` can be used which not only supports binary but also polytomous item responses and they have corresponding methods for printing, plotting, subsetting, etc., some of which are illustrated below.

In addition to the item responses, there are a number of covariates of interest:

- `group`: Factor for group (1 vs. 2).
- `tests`: Number of previous online exercises solved (out of 26).
- `nsolved`: Number of exam items solved (out of 13).
- `gender`, `study`, `attempt`, `semester`.

For a first overview, we load the package and data. Then we exclude those participants with the extreme scores of 0 or 13, respectively, because these students do not discriminate between the items (either none solved or all solved). The R code below employs the `print()` and `plot()` methods for '`itemresp`' objects by printing the first couple of item responses and visualizing the proportion of correct responses per item.

```
R> library("psychotools")
R> data("MathExam14W", package = "psychotools")
R> mex <- subset(MathExam14W, nsolved > 0 & nsolved < 13)
R> head(mex$solved)
```

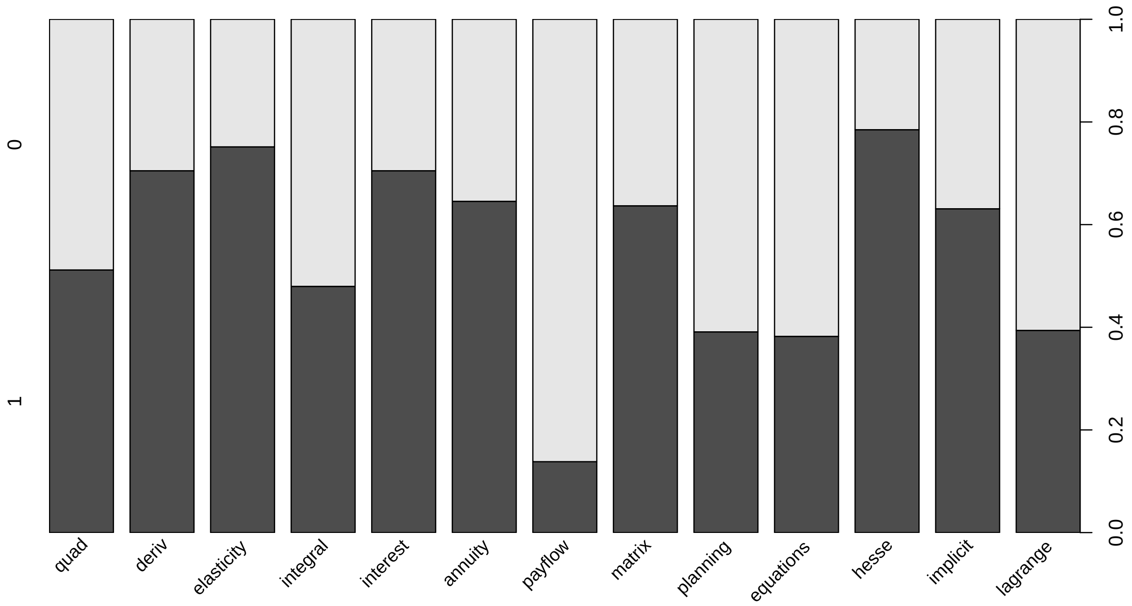


Figure 1: Bar charts with relative frequencies of items solved correctly (1, dark gray) or not solved correctly (0, light gray). The item labels on the x-axis briefly indicate the topic of each item.

```
[1] {1,1,1,0,1,1,0,1,1,0,1,1,0} {1,1,1,1,0,0,0,1,1,1,1,1,1}
[3] {0,0,1,0,0,1,0,1,1,0,0,1,0} {0,1,0,1,1,1,1,1,1,1,1,1,1}
[5] {1,0,0,1,1,0,0,0,0,1,1,0,1} {1,0,0,1,1,0,0,0,1,1,0,0,0}
```

```
R> plot(mex$solved)
```

The resulting plot is shown in Figure 1, highlighting that most items have been solved correctly by about 40 to 80 percent of the participants. The main exception is the payflow exercise (for which a certain integral had to be computed) which was solved correctly by less than 15 percent of the students.

3. IRT with the Rasch model

While the exploratory analysis of the item responses is already interesting, item response theory (IRT, Fischer and Molenaar 1995; Van der Linden 2016) provides us with more refined methods to infer the underlying latent trait captured by the exam (here: ability in basic mathematics). In particular, it allows to align the items' difficulties and the students' abilities on the same scale for the latent trait.

Here, we employ the Rasch model (Rasch 1960) which is widely used in educational and psychological assessments, e.g., in the PISA programme (OECD 2024). The Rasch model aligns the ability θ_i of person i ($i = 1, \dots, n$) and the difficulty β_j of item j ($j = 1, \dots, m$) using a logistic model for the observed binary items $\{y_{ij}\} \in \{0, 1\}$ from all n persons and m items.

$$\begin{aligned} \pi_{ij} &= \Pr(y_{ij} = 1) \\ \mathrm{logit}(\pi_{ij}) &= \theta_i - \beta_j \end{aligned}$$

This means that the probability π_{ij} that person i solves item j correctly is linked to the difference of ability and difficulty through a logit link where $\mathrm{logit}(\pi) = \log\{\pi/(1 - \pi)\}$.

Most statistical audiences will be familiar with fitting and interpreting logistic regression models for binary responses, but there are two aspects in the Rasch model that require special attention: First, full maximum likelihood estimation of all model parameters (also known as joint maximum likelihood) is inconsistent because the number of parameters also diverges as the number of observations increases to infinity (either by increasing the number of persons n or the number of items m). Second, the model parameters θ_i and β_j are only defined up to an additive constant because only their difference is identified and adding the same constant to both parameter groups always cancels out.

The first issue is usually solved by one of two approaches: Either some (typically normal) distribution is assumed for the ability parameters which can then be leveraged for the item parameter estimation – using marginal maximum likelihood in a frequentist setting or as a prior distribution in Bayesian estimation. Alternatively, the abilities for person i are conditioned on the so-called sum scores (the number of items solved by that person) and conditional maximum likelihood is used (Fischer and Molenaar 1995). Throughout this paper we employ conditional maximum likelihood estimation. See von Davier (2016) for more details and further estimation techniques.

The second issue is solved by fixing a zero point on the latent trait scale, e.g., by fixing the sum of all item parameters to zero. As long as only a single model is fitted to all persons, the specific reference point is usually not of much practical relevance. However, it is crucial when comparing the item parameter estimates from two or more subgroups of persons. This will be revisited in the next section. For now, we restrict the sum of item parameters to zero.

In R a number of packages are available to estimate various IRT models, including the Rasch model (see Debelak, Strobl, and Zeigenfuse 2022, for an introduction). The most popular ones are probably **eRm** (Mair and Hatzinger 2007) and **mirt** (Chalmers 2012). Here, we employ the implementation in package **psychotools** (Zeileis *et al.* 2024a) instead because it provides a particularly rich toolbox of methods for assessing measurement invariance which we leverage in the next section.

Fitting a Rasch model with **psychotools** can be carried out with the function `raschmodel()` based on a binary item response matrix. For the resulting model class a number of methods are available, e.g., extracting the item and person parameters via `itempar()` and `personpar()`, respectively. These parameters can then be displayed using the `plot()` methods in various ways, e.g., just the "profile" of item parameters (Figure 2, top) or a so-called person-item plot (Figure 2, bottom) which displays both sets of parameters on the same scale.

```
R> mr <- raschmodel(mex$solved)
R> plot(mr, type = "profile")
R> plot(mr, type = "piplot")
```

Qualitatively, both model-based displays in Figure 2 show a similar pattern as the empirical proportions from Figure 1. However, due to the latent logistic scale the most difficult item (payflow) and the easiest item (hesse) are brought out even more clearly. Also, the person-item plots conveys that the majority of the item difficulties are close to the median ability in this sample. Thus, the exam discriminates more sharply at the median difficulty and less sharply in the tails at very high or very low ability.

4. Assessment of measurement invariance

The interpretation of the Rasch model parameters from the previous section is only reliable if all item difficulties β_j are indeed the same for all students in the sample. If this is not the case, differences in the item responses would not necessarily be caused by differences in basic mathematics ability. The fundamental assumption that the item parameters are constant across all persons is a special case of so-called measurement invariance. And a violation of this assumption is known as differential item functioning (DIF, Gamerman, Gonçalves, and

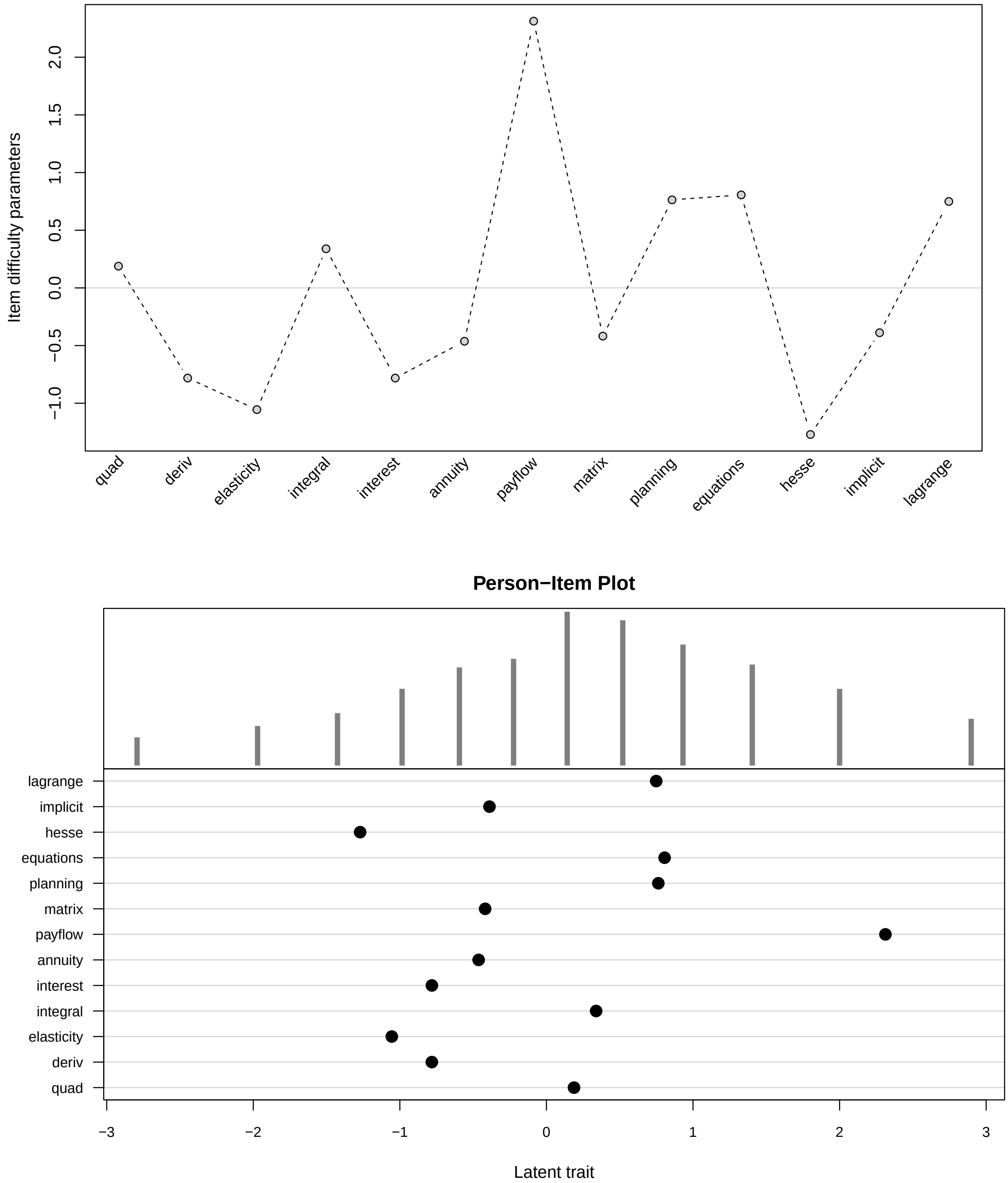


Figure 2: Top: Profile plot of estimated item parameters (difficulties $\hat{\beta}_j$). Bottom: Person-item plot of estimated person parameters (abilities $\hat{\theta}_i$, bar plot in upper panel) and item parameters (difficulties $\hat{\beta}_j$, dot plot in lower panel).

Soares 2016; Debelak and Strobl 2024), i.e., some item(s) is/are relatively easier for some subgroup of persons compared to others.

Many techniques have been proposed in the literature for detecting DIF, including nonparametric tests, graphical methods, or regression-based approaches (see, e.g., Gamerman *et al.* 2016; Magis, Béland, Tuerlinckx, and De Boeck 2010; Hladká and Martinková 2020, among others). In the following we focus on methods that are based on parametric psychometric models, where a violation of measurement invariance implies that the item parameters differ across persons (Merkle and Zeileis 2013), e.g., along with some available covariates. For example, item difficulty might vary along with age or some items might be more difficult for non-native speakers etc. Note that this is *not* the situation where a person's ability varies along with some covariates because this would imply that for such a person *all* items are relatively more or less difficult. In contrast, DIF refers to the situation where this occurs only for some item(s).

In the following, we consider three different methods for assessing DIF in the Rasch model: First, if the potential subgroups affected by DIF are known, it is possible to test for item parameter differences between these groups using the classical likelihood ratio, score, or Wald tests with suitable item anchoring (Glas and Verhelst 1995). Second, generalizations of the score test can also assess changes along continuous or ordinal variables (Merkle and Zeileis 2013; Merkle, Fan, and Zeileis 2014) and trees can repeat these tests recursively for a combination of different variables (nominal, ordinal, or continuous) to form subgroups in a data-driven way (Strobl, Kopf, and Zeileis 2015). Third, finite mixtures of Rasch models are a general strategy to test for violations of measurement invariance, even when there are no observed covariates (originally proposed by Rost 1990). But it is also possible to inform the selection of subgroups in Rasch mixture models by covariates (see Frick, Strobl, Leisch, and Zeileis 2012).

4.1. Classical tests with reference and focal groups

In the examination of our mathematics exam results, the obvious first question in a DIF setting is whether the item parameters for the first and the second group differ. Recall that the two groups took the exam in back to back sessions and the item categories were the same for all 13 items. However, different concrete items were used from these categories for the first and the second group. Thus, it is natural to suspect that at least some of the items might have different difficulties.

The strategy for this is straightforward: We split the data into reference and focal groups (here `group` 1 and 2) and then assess the stability of selected parameters across the groups by means of standard tests. Especially, the likelihood ratio (LR) test is easy to compute. Its test statistic is twice the difference between the full-sample likelihood and the overall likelihood from the two subgroups. Using the **psychotools** functionality the LR statistics can be obtained "by hand" via:

```
R> mr1 <- raschmodel(subset(mex, group == 1)$solved)
R> mr2 <- raschmodel(subset(mex, group == 2)$solved)
R> -2 * as.numeric(logLik(mr) - (logLik(mr1) + logLik(mr2)))
```

```
[1] 264.9577
```

This shows that the conditional likelihood of the model can be improved a lot by splitting into `group` 1 and 2 and that the LR test statistic is much larger than the 95% critical value 21.0 from the χ^2_{12} null distribution (because the split necessitates the estimation of 12 additional item parameters). Very similar results would be obtained when using the Wald statistic 249.4 or the score (or Lagrange multiplier) statistic 260.8.

Given that there is such strong evidence for DIF between the two groups, the natural next question is: Which items "cause" this DIF? A natural strategy for answering this question is looking at the item-wise Wald tests. This is simply the difference between the item parameter estimate for item j from the reference and the focal group, scaled by the corresponding standard error:

$$t_j = \frac{\hat{\beta}_j^{\text{ref}} - \hat{\beta}_j^{\text{foc}}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}^{\text{ref}})_{j,j} + \widehat{\text{Var}}(\hat{\beta}^{\text{foc}})_{j,j}}}.$$

However, there is one important caveat: As the reference point of zero from the two scales is arbitrary, we need to find a way to "anchor" the scales from the two groups. Typically, this is done by selecting either a single item or groups of items whose (mean of) parameters are used as the zero reference point (Woods 2009). Unfortunately, though, the selection of the anchor item(s) has a large influence on the results of the item-wise Wald tests and there is no obvious selection strategy. This is illustrated in Figure 3 which shows anchoring with item 1 (quad, top panel) and item 10 (equations, bottom panel). The top panel is essentially generated as follows:

```
R> plot(mr1, parg = list(ref = 1), col = 2, ylim = c(-2.6, 2.6))
R> plot(mr2, parg = list(ref = 1), col = 4, add = TRUE)
```

The bottom panel analogously uses `ref = 10`.

With the anchoring in the top panel one would have to conclude that almost all items differ in their difficulties except the fixed item 1 (quad) and item 9 (planning). However, a completely different picture emerges for the anchoring with item 10 in the bottom panel. Note that the patterns within each profile are exactly the same, they are just both shifted vertically so that the item parameter for item 10 is zero in both groups. With this anchoring, only items 1 (quad), 7 (payflow), and 9 (planning) differ more substantially whereas the remaining items function very similarly.

The latter interpretation is much more plausible and can also be backed up by looking at the content of item 10. This was the same item template across both groups, namely solving a system of three linear equations, where all students just had somewhat different coefficients within the equations. Thus, this item is a natural candidate for being an anchor item as it can be expected to function the same for both groups.

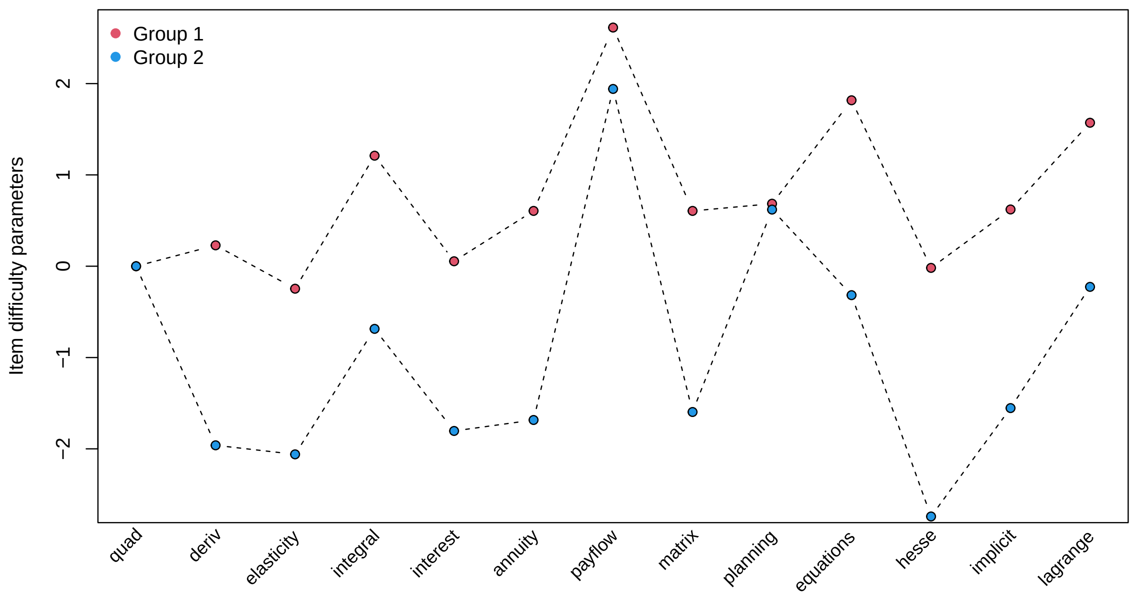
To further validate this conclusion, we also consider a data-driven choice of the anchor item(s). A lot of different strategies have been proposed for this purpose (for an overview see Kopf, Zeileis, and Strobl 2015a,b) which are based on the basic idea to select some DIF-free anchor item(s) to be able to identify items truly associated with DIF. Given that this is a kind of "the chicken or the egg" dilemma, it is not so easy to solve and there is no consensus as to what is the best way to do so. Here, we use the recently proposed strategy by Strobl, Kopf, Kohler, von Oertzen, and Zeileis (2021), which has been shown to have a simple motivation and good empirical performance. This strategy selects the anchor item that minimizes the inequality (as measured by the so-called Gini coefficient) among the item parameter differences. Here, this yields item 12 (implicit) as the best anchor item. However, the results are very similar to those from using item 10, leading to qualitatively the same insights.

In **psychotools** this anchor strategy is the default in the `anchortest()` function. We call it and indicate that we want to test for DIF in the `solved` items by `group` and adjust for multiple testing (across the 12 estimated item parameters) with the single-step method from **multcomp** (Hothorn, Bretz, and Westfall 2008).

```
R> ma <- anchortest(solved ~ group, data = mex, adjust = "single-step")
R> plot(ma)
```

The visualization from the `plot()` method (see Figure 4) does not show the two item profiles (as in Figure 3) but shows confidence intervals for all item parameter differences (except the anchor item whose difference is fixed to zero). This brings out clearly that items 1 (quad), 7 (payflow), and 9 (planning) are significantly more difficult for `group` 2 than for `group` 1, relative to the remaining items which do not significantly violate the measurement invariance assumption.

As there is significant evidence for DIF, it is clear that the results for `group` 1 and 2 cannot be compared completely fairly. However, as the students were allowed to select the earlier vs. later group themselves up until about a week before the exam (when the remaining students



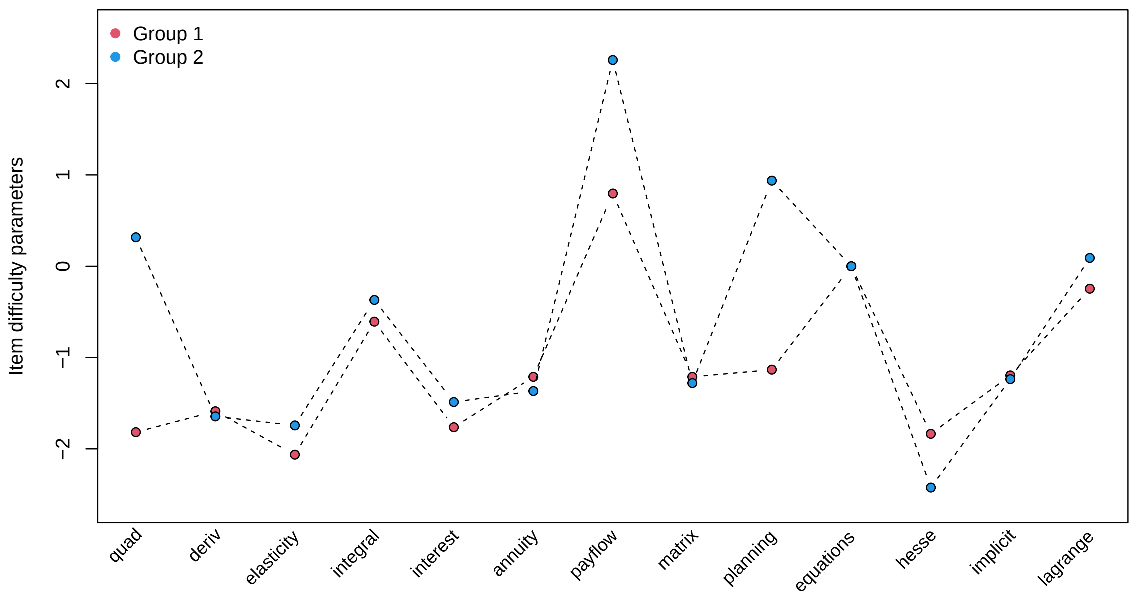


Figure 3: Top: Profile plot of estimated item parameters (difficulties $\hat{\beta}_j$) with the constraint $\hat{\beta}_1 = 0$. Bottom: Profile plot with the constraint $\hat{\beta}_{10} = 0$.

were assigned randomly), it cannot be ruled out that the students in `group` 1 and 2 differ systematically in some way which in turn could have lead to the difference in difficulties. It is more likely, though, that the main source for the differences is that slightly different concrete items were used for the two groups. Hence, `group` 2 appears to have been somewhat disadvantaged by getting the more difficult version of the exercises. Moreover, the question remains if there is even further DIF within these two groups which might be associated with some of the other available covariates.

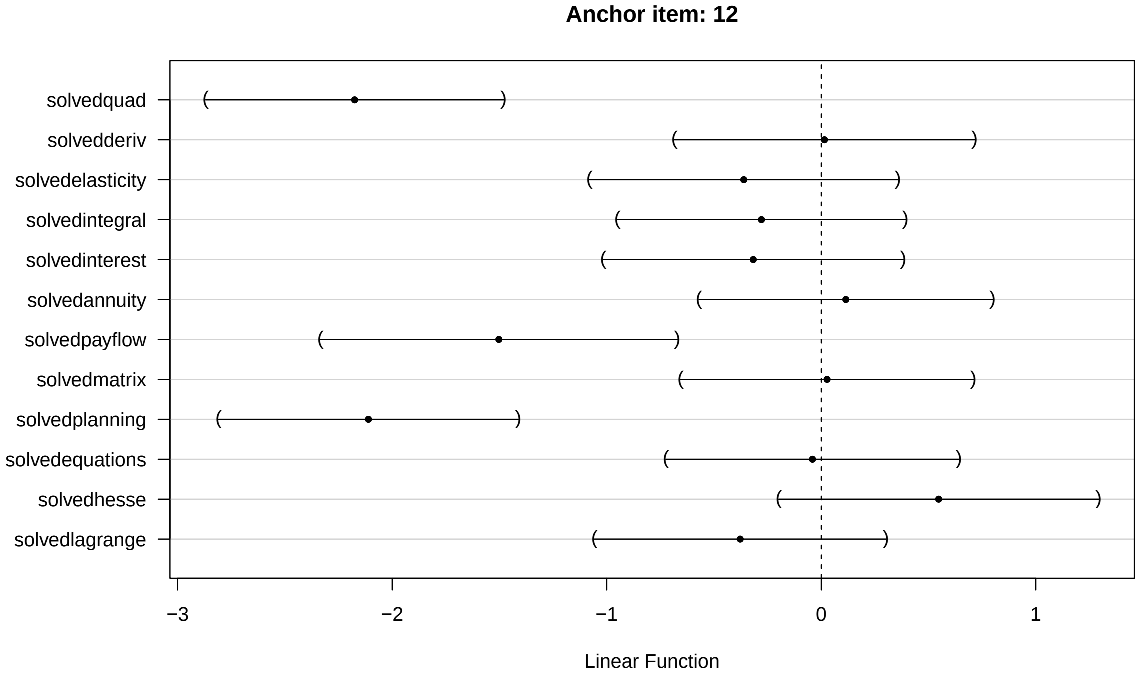


Figure 4: Point estimates and corresponding simultaneous confidence intervals for all item parameter differences between group 1 minus group 2, except for anchor item 12 (implicit) whose item parameter estimates are restricted to zero in both groups.

4.2. Rasch trees

The comparison between a reference and a focal group using classical tests is very nice and straightforward for those familiar with maximum likelihood inference. However, the main drawback is that the potential subgroups have to be formed in advance. This is particularly inconvenient when testing for DIF along continuous variables which, in practice, are often categorized into groups in an ad hoc way (e.g., splitting at the median). Also, when considering DIF in ordinal variables, the ordering of categories is often not exploited, thus assessing only if at least one group differs from the others.

To address these problems, Merkle and Zeileis (2013) and Merkle *et al.* (2014) have proposed generalizations of the classical score test which allow to assess parameter instabilities along numeric and ordinal covariates, respectively. These tests require no split into subgroups but just rely on the ordering implied by the covariate.

Here, we first illustrate one of their tests for a numeric covariate by testing for DIF along `tests` in `group` 1. Recall that `tests` is the number of points obtained in the online tests throughout the semester and thus captures a certain skill level and amount of preparation prior to the exam. The test statistic used is simply the maximum of the score (or Lagrange multiplier) statistics for each possible split in `tests`. Figure 5 shows the sequence of score statistics beginning with the one where students with `tests` ≤ 9 form the reference group and `tests` > 9 the focal group. Then the same kind of statistics are computed for splits at 11 points in the `tests`, for 13, etc. Eventually, the null hypothesis has to be rejected if the maximum of these tests becomes larger than the corresponding critical value, which, of course, has to be adjusted for the fact that the maximum of multiple test statistics is considered here. Figure 5 shows the 95 percent critical value as the horizontal red line which is clearly exceeded by the sequence of test statistics. The highest test statistics, and thus the highest amount of DIF, is obtained for a split at `tests` $= 16$.

In R, this maximum score test can be carried out using the **strucchange** package (Zeileis, Leisch, Hornik, and Kleiber 2002) which produces Figure 5 and also obtains a p-value by

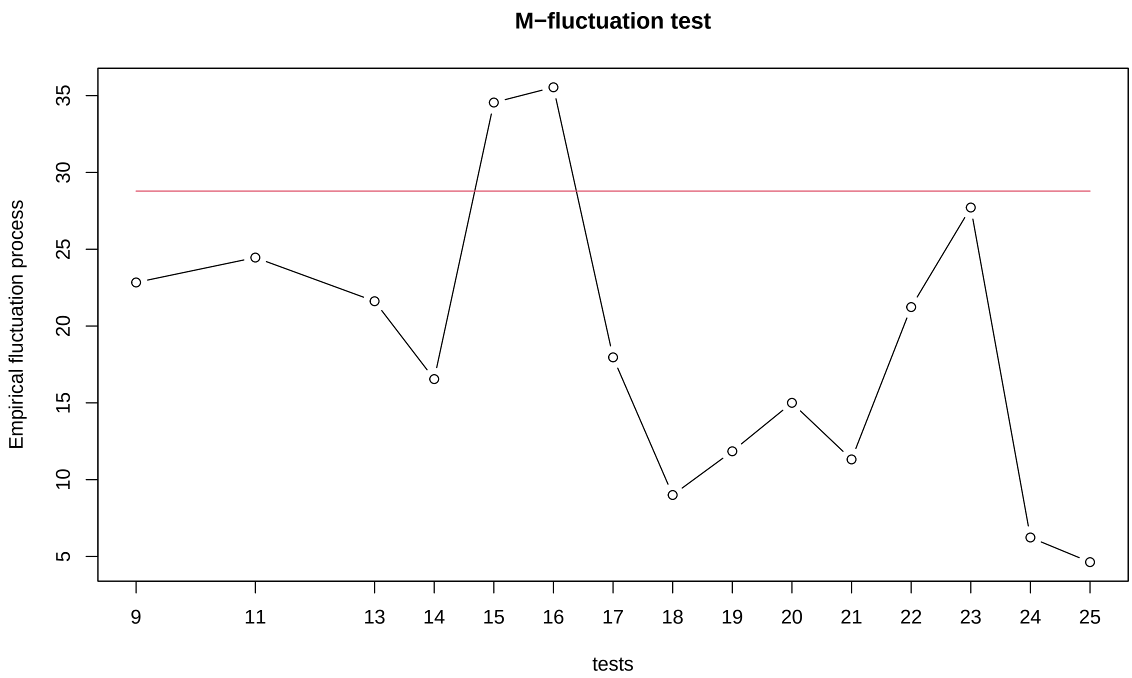


Figure 5: Sequence of score (or Lagrange multiplier) test statistics for splits in the number of points in the online tests (within the first group). The horizontal red line is the 95 percent critical value for the maximum of the score statistics.

simulation.

```
R> library("strucchange")
R> mex1 <- subset(mex, group == 1)
R> sctest(mr1, order.by = mex1$tests, vcov = "info", functional = "maxLMo",
+    plot = TRUE)

        M-fluctuation test

data:  mr1
f(efp) = 35.543, p-value = 0.005373
```

The result indicates that within the first group there is further evidence for DIF. Students who performed rather poorly in the previous online tests have a different item profile. More details will be provided below.

Now the natural next step is again to form subgroups for `tests` ≤ 16 and `tests` > 16, respectively, inspect the corresponding item parameter profiles, and possibly test for further DIF. This strategy can be formalized using so-called model-based recursive partitioning (Zeileis, Hothorn, and Hornik 2008) in combination with Rasch models. The resulting Rasch trees (Strobl *et al.* 2015) are constructed recursively in the following steps:

- Fit a Rasch model to the current subsample.
- Assess DIF along all covariates of interest (applying a Bonferroni adjustment for multiple testing).
- Split with respect to the covariate with the smallest significant p-value.
- Select split point by maximizing the log-likelihood.

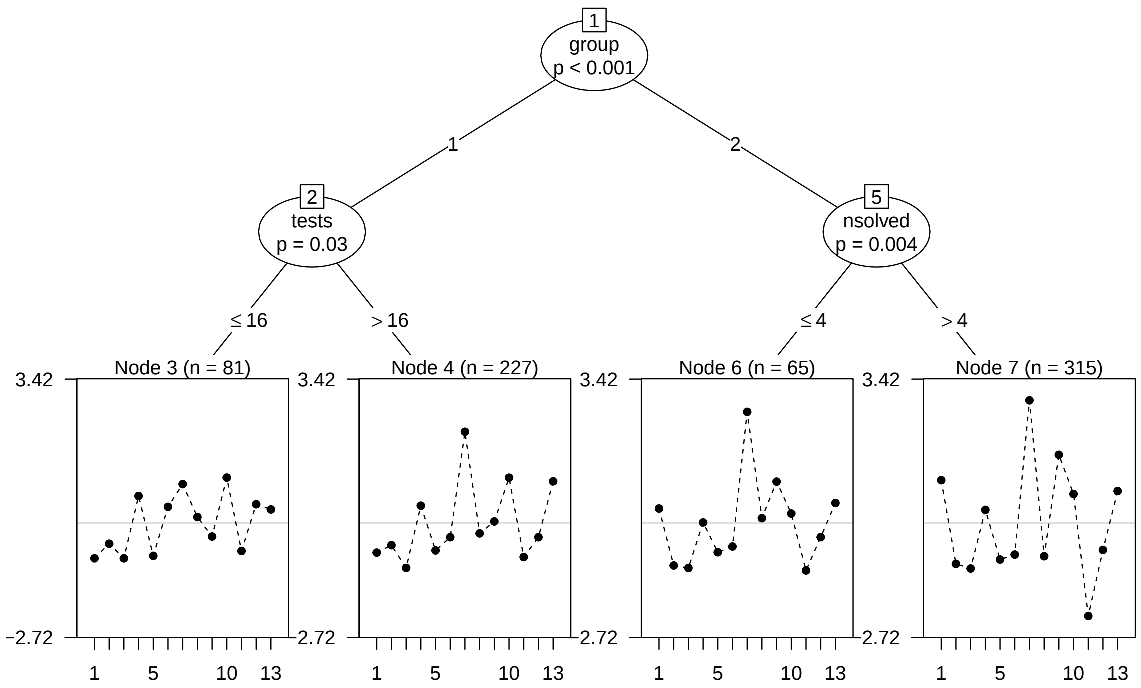


Figure 6: Rasch tree for the mathematics exam data, detecting DIF with respect to the exam group (1 vs. 2) and with respect to the mathematical abilities, captured by the number of points in the online tests and the number of items solved in the exam, respectively.

- Continue until there are no more significant instabilities (or the sample is too small).

This strategy can be applied relatively easily using the **psychotree** package which combines the model-based recursive partitioning infrastructure from **partykit** (Hothorn and Zeileis 2015) with the IRT models from **psychotools**. To apply the method to the mathematics exam data, we treat all numeric variables as ordinal because they have relatively few distinct levels.

```
R> library("psychotree")
R> mex <- transform(mex,
+    tests    = ordered(tests),
+    nsolved  = ordered(nsolved),
+    attempt  = ordered(attempt),
+    semester = ordered(semester)
+  )
R> mrt <- raschtree(solved ~ group + tests + nsolved + gender +
+    attempt + study + semester, data = mex,
+    vcov = "info", minsize = 50, ordinal = "L2", nrep = 1e5)
```

The arguments in the last line specify that the information matrix is used as the estimate for the variance-covariance matrix, that each subgroup must have at least 50 persons, and that the maximum score test is used for the ordinal covariates (which had already been used in the `sctest()` above where the corresponding argument was called `functional = "maxLMo"`) with `nrep` replications in the simulated p-values.

The resulting tree can be visualized with `plot(mrt)` and is depicted in Figure 6. The tree detects all of the subgroups we had considered above in a data-driven way. First, the highest amount of DIF is found with respect to the first vs. second group. The differences can be seen between the two item profile plots on the left (nodes 3 and 4) and those on the right (nodes 6 and 7). For example, in the second group the first item was relatively much more difficult than the subsequent items while it was somewhat easier in the first group. Similarly, item 9

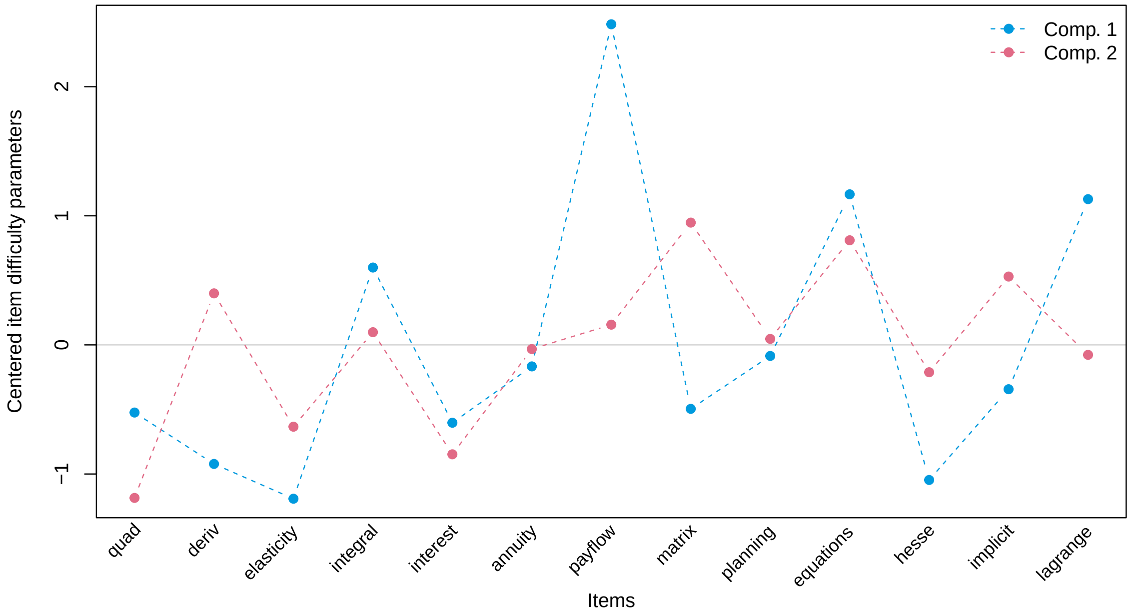


Figure 7: Item profile plots for 2-component Rasch mixture model for the students in group 1.

appears to be easier than item 10 in the first group but more difficult in the second group. Within both groups further evidence for DIF is found so that they are split into a smaller subgroup of persons with lower mathematical abilities (nodes 3 and 6, respectively) and a larger subgroup of persons with higher mathematical abilities (nodes 4 and 7, respectively). In the first group this split is characterized by the number of points obtained in the online tests during the semester while in the second group it is with respect to the number of items solved correctly within the exam. Qualitatively, the change in the item profiles is similar within both groups. For the students with higher mathematical abilities there is a clear profile that distinguishes between easier and harder items. In contrast, for the students with lower mathematics skills, the item profile is very much "dampened" (especially in group 1), reflecting that for them all items are almost equally easy or hard.

4.3. Rasch mixture models

Finally, we consider another possibility to assess measurement invariance in Rasch models when there are no covariates available at all. This can be accomplished using finite mixtures of Rasch models (Rost 1990). The idea is that there are two or more subgroups or clusters with different item parameters and that the cluster membership is *not* known. But it can be inferred using the EM (expectation-maximization) algorithm which iterates between estimating the parameters in all clusters (M-step) and then determining the a-posteriori probabilities of each observation to belong to each of the clusters (E-step).

In R, the package **flexmix** (Leisch 2004; Grün and Leisch 2008) provides a modular implementation of finite mixture models which Frick *et al.* (2012) have interfaced in **psychomix** for IRT modeling with the **psychotools** package. Frick, Strobl, and Zeileis (2015) further extended the functionality by considering enhancements for selecting the necessary number of clusters and for appropriately modeling the cluster-specific distribution of sum scores.

Here, we only illustrate the method briefly by asking whether it would be possible to detect DIF in `group` 1 without knowing the relevant covariate `tests`. We do so by directly fitting a 2-component mixture (without selecting the number of clusters) with component-specific sum score distribution in mean-variance specification.

```
R> library("psychomix")
R> mrm <- raschmix(mex1$solved, k = 2, scores = "meanvar")
R> plot(mrm)
R> print(mrm)

Call:
raschmix(formula = mex1$solved, k = 2, scores = "meanvar")

Cluster sizes:
  1   2
235  73

convergence after 73 iterations
```

The result of the Rasch mixture model is described by the `print()` output above and the `plot()` output in Figure 7. Qualitatively, this provides the same insights as nodes 3 and 4 in Figure 6. There is a larger group that clearly discriminates between the items that are relatively easy and those that are relatively difficult. But there is also a smaller group whose item profile is very much "dampened" and for whom all items are almost equally easy or equally difficult.

The main difference is that here the Rasch tree is better at picking up the subgroups because it can exploit the available covariates. Moreover, the tree yields "hard" splits where a person is only in exactly one subgroup of the tree. In contrast, in the finite mixture model, there are only "soft" classifications so that each person is part of all clusters but with different weights. See Frick, Strobl, and Zeileis (2014) for a comparison of regression trees and finite mixtures and what their relative advantages and disadvantages are.

5. Discussion

This paper presents a hands-on tutorial for examining the results from multiple-choice exams using IRT models along with methods for assessing measurement invariance, in particular differential item functioning (DIF). The flexible toolbox encompasses methods for detecting violations along one covariate (tests), many covariates (trees), or no covariates (mixtures). The covariates do not need to be dichotomized in advance but can exploit their different scales, be it continuous, ordinal, or categorical.

For the data set presented, results from an introductory mathematics exam, along with a number of different covariates, the Rasch tree gives probably the quickest overview of the underlying DIF patterns without having to specify specific subgroups in advance.

The insights from this analysis had a number of policy implications for the introductory mathematics exams at Universität Innsbruck. First, exam groups are avoided, if at all possible. Second, seemingly equivalent items can function very differently if students focus their learning on well-known parts of the item pool. This was the case for the first item (on optimization of a quadratic function). Both question texts were very similar in terms of the mathematical problem they described. However, one story was used in the online tests during the same semester (group 2) and the other story was used in the end-term exam from the previous semester (group 1) that all students had access to. And it turned out that the students perceived the latter exercise as much easier, possibly because they practiced with last semester's exam more than with the online test exercises.

The analysis ties together a number of R packages, all of which are freely available under the General Public License from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/`. **strucchange** provides an object-oriented implementation of the score-based parameter instability tests. Model-based recursive partitioning is available in **partykit** and model-based clustering with finite mixture models in **flexmix**. Psychometric models that cooperate with **strucchange**, **partykit**, and **flexmix**, are implemented in **psychotools**, including various IRT models (Rasch, partial credit, rating scale, parametric logistic), Bradley-Terry, and multinomial processing tree models. The corresponding psychometric trees are in **psychotree** and the psychometric mixture models in **psychomix**.

Thus, using the packages above it would also be possible to employ more refined IRT models in the DIF analysis, e.g., logistic models with two or more parameters. Moreover, the analysis could be complemented or extended by further DIF detection methods, both based on IRT models and beyond (e.g., Magis *et al.* 2010; Hladká and Martinková 2020). See also Mair (2018) and Martinková and Hladká (2023) for further psychometric techniques available in R.

Finally, the R package **exams** provides flexible infrastructure for conducting large-scale exams using randomized dynamic item pools either in classical written exams (optionally with automatic evaluation) or in online learning management systems (such as Moodle, OpenOlat, Canvas, Blackboard, etc.). It is based on exercise templates with single- or multiple-choice items, numeric exercises, closed and open-ended text questions, as well as combinations of all of these.

6. Epilogue

This paper is dedicated to the memory of our friend and colleague Friedrich "Fritz" Leisch who died after serious illness in April 2024. Fritz contributed to the work presented here in a number of ways: He co-developed the **strucchange** package and some of the parameter instability tests which were originally geared towards testing for structural changes in time series regressions. And it was only somewhat later that we realized that we could adapt the same tests for recursive partitioning in regression trees (Zeileis *et al.* 2008). While Fritz was not a co-author of the latter work, he had nevertheless been an influence for it, mainly due to his work on model-based clustering with finite mixture models (Leisch 2004). The way that Fritz had set up a modular framework for plugging different kinds of models into **flexmix** inspired us to establish a similar model-based framework for recursive partitioning along with an object-oriented implementation. Later on we came full circle by developing together the Rasch mixture models in **psychomix** (Frick *et al.* 2012). Finally, Fritz was also a co-author of the R/**exams** infrastructure. He was the first who had successfully extended our original implementation (which was geared towards PDF output only) to XML exports for the Moodle learning management system. So we joined forces to establish a flexible toolbox that can generate all kinds of different exports from the same dynamic item pool.

Fritz will be dearly missed as scholar, collaborator, and friend.

Acknowledgments

Helpful feedback from the guest editors and two reviewers is gratefully acknowledged. This substantially improved the paper – all remaining errors are our own.

References

Anthology Inc (2024). *Blackboard 3900.100.* URL `https://www.anthology.com/`.

Chalmers RP (2012). "**mirt**: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software*, **48**(6), 1–29. `doi:10.18637/jss.v048.i06`.

Debelak R, Strobl C (2024). "Violations of Unidimensionality and Differential Item Functioning." In S Greiff, K Schweizer, S Troche (eds.), *Method Effects in the Psychological Measurement.* Hogrefe.

Debelak R, Strobl C, Zeigenfuse MD (2022). *An Introduction to the Rasch Model with Examples in R*. Chapman & Hall/CRC, New York. doi:10.1201/9781315200620.

Dougiamas M, *et al.* (2024). *Moodle, Version 4.5*. URL https://moodle.org/.

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

frentix GmbH (2024). *OpenOlat 19.0: Open Source Learning Management System*. URL https://www.openolat.com/.

Frick H, Leisch F, Strobl C, Wickelmaier F, Zeileis A (2024). ***psychomix**: Psychometric Mixture Models*. doi:10.32614/CRAN.package.psychomix. R package version 1.1-9.

Frick H, Strobl C, Leisch F, Zeileis A (2012). "Flexible Rasch Mixture Models with Package **psychomix**." *Journal of Statistical Software*, **48**(7), 1–25. doi:10.18637/jss.v048.i07.

Frick H, Strobl C, Zeileis A (2014). "To Split or to Mix? Tree vs. Mixture Models for Detecting Subgroups." In M Gilli, G González-Rodríguez, A Nieto-Reyes (eds.), *Proceedings of COMPSTAT 2014 – 21st International Conference on Computational Statistics*, pp. 379–386.

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.

Gamerman D, Gonçalves FB, Soares TM (2016). "Differential Item Functioning." In WJ Van der Linden (ed.), *Handbook of Item Response Theory: Volume Three*, chapter 4. Chapman & Hall/CRC, New York.

Glas CAW, Verhelst ND (1995). "Testing the Rasch Model." In GH Fischer, IW Molenaar (eds.), *Rasch Models: Foundations, Recent Developements, and Applications*, chapter 5. Springer-Verlag, New York.

Grün B, Leisch F (2008). "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software*, **28**(4), 1–35. doi:10.18637/jss.v028.i04.

Grün B, Zeileis A (2009). "Automatic Generation of Exams in R." *Journal of Statistical Software*, **29**(10), 1–14. doi:10.18637/jss.v029.i10.

Hladká A, Martinková P (2020). "**difNLR**: Generalized Logistic Regression Models for DIF and DDF Detection." *The R Journal*, **12**(1), 300–323. doi:10.32614/rj-2020-014.

Horn JL, McArdle JJ (1992). "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research*, **18**(3), 117–144. doi:10.1080/03610739208253916.

Hothorn T, Bretz F, Westfall P (2008). "Simultaneous Inference in General Parametric Models." *Biometrical Journal*, **50**(3), 346–363. doi:10.1002/bimj.200810425.

Hothorn T, Zeileis A (2015). "**partykit**: A Modular Toolkit for Recursive Partytioning in R." *Journal of Machine Learning Research*, **16**, 3905–3909.

Instructure, Inc (2024). *Canvas*. URL https://www.instructure.com/.

Kopf J, Zeileis A, Strobl C (2015a). "Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches." *Educational and Psychological Measurement*, **75**(1), 22–56. doi:10.1177/0013164414529792.

Kopf J, Zeileis A, Strobl C (2015b). "A Framework for Anchor Methods and an Iterative Forward Approach for DIF Detection." *Applied Psychological Measurement*, **39**(2), 83–103. doi:10.1177/0146621614544195.

Leisch F (2004). "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R." *Journal of Statistical Software*, **11**(8), 1–18. doi:10.18637/jss.v011.i08.

Magis D, Béland S, Tuerlinckx F, De Boeck P (2010). "A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning." *Behavior Research Methods*, **42**(1), 847–862. doi:10.3758/brm.42.3.847.

Mair P (2018). *Modern Psychometrics with R*. Springer-Verlag, Cham. doi:10.1007/978-3-319-93177-7.

Mair P, Hatzinger R (2007). "Extended Rasch Modeling: The **eRm** Package for the Application of IRT Models in R." *Journal of Statistical Software*, **20**(9), 1–20. doi:10.18637/jss.v020.i09.

Martinková P, Hladká A (2023). *Computational Aspects of Psychometric Methods*. Chapman & Hall/CRC, New York. doi:10.1201/9781003054313.

Merkle EC, Fan J, Zeileis A (2014). "Testing for Measurement Invariance with Respect to an Ordinal Variable." *Psychometrika*, **79**(4), 569–584. doi:10.1007/S11336-013-9376-7.

Merkle EC, Zeileis A (2013). "Tests of Measurement Invariance without Subgroups: A Generalization of Classical Methods." *Psychometrika*, **78**(1), 59–82. doi:10.1007/s11336-012-9302-4.

Nettekoven M, Ledermüller K (2012). "Assess the Assessment: An Automated Analysis of Multiple Choice Exams and Test Items." In H Beldhuis (ed.), *Proceedings of the 11th European Conference on E-Learning*, pp. 397–406.

Nettekoven M, Ledermüller K (2014). "MC Monitoring: Automated Evaluation of Multiple Choice Exams and Test Items." In International Association of Technology, Education and Development (IATED) (ed.), *Proceedings of the 8th International Technology, Education and Development Conference*, pp. 4023–4032.

OECD (2024). *PISA 2022 Results (Volume V): Learning Strategies and Attitudes for Life*. OECD Publishing, Paris. doi:10.1787/c2e44201-en.

Rasch G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedogogiske Institut, Kopenhagen.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rost J (1990). "Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis." *Applied Psychological Measurement*, **14**(3), 271–282. doi:10.1177/014662169001400305.

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**(2), 289–316. doi:10.1007/s11336-013-9388-3.

Van der Linden WJ (ed.) (2016). *Handbook of Item Response Theory: Three Volume Set*. Chapman & Hall/CRC, New York. doi:10.1201/9781315119144.

von Davier M (2016). "Rasch Model." In WJ Van der Linden (ed.), *Handbook of Item Response Theory: Volume One*, chapter 3. Chapman & Hall/CRC, New York.

Wikipedia (2024). "Learning Analytics — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/Learning_analytics, accessed 2024-12-30.

Woods CM (2009). "Empirical Selection of Anchors for Tests of Differential Item Functioning." *Applied Psychological Measurement*, **33**(1), 42–57. doi:10.1177/0146621607314044.

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, Leisch F, Hornik K, Kleiber C (2002). "**strucchange**: An R Package for Testing for Structural Change in Linear Regression Models." *Journal of Statistical Software*, **7**(2), 1–38. doi:10.18637/jss.v007.i02.

Zeileis A, Strobl C, Wickelmaier F, Komboz B, Kopf J, Schneider L, Debelak R (2024a). ***psychotools**: Infrastructure for Psychometric Modeling*. doi:10.32614/CRAN.package.psychotools. R package version 0.7-4.

Zeileis A, Strobl C, Wickelmaier F, Komboz B, Kopf J, Schneider L, Dreifuss D, Debelak R (2024b). ***psychotree**: Recursive Partitioning Based on Psychometric Models*. doi:10.32614/CRAN.package.psychotree. R package version 0.16-1.

Zeileis A, Umlauf N, Leisch F (2014). "Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond." *Journal of Statistical Software*, **58**(1), 1–36. doi:10.18637/jss.v058.i01.

Affiliation:

Achim Zeileis
Universität Innsbruck
Department of Statistics
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org,
URL: https://www.zeileis.org/

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 54 *Submitted:* 2024-09-29
2025 *Accepted:* 2025-02-04