Rasch Mixture Models for DIF Detection: A Comparison of Old and New Score Specifications

Hannah Frick
Universität Innsbruck

Carolin Strobl
Universität Zürich

Achim Zeileis
Universität Innsbruck

Abstract

This vignette is a (slightly) modified version of Frick, Strobl, and Zeileis (2014), published in *Educational and Psychological Measurement*.

Rasch mixture models can be a useful tool when checking the assumption of measurement invariance for a single Rasch model. They provide advantages compared to manifest DIF tests when the DIF groups are only weakly correlated with the manifest covariates available. Unlike in single Rasch models, estimation of Rasch mixture models is sensitive to the specification of the ability distribution even when the conditional maximum likelihood approach is used. It is demonstrated in a simulation study how differences in ability can influence the latent classes of a Rasch mixture model. If the aim is only DIF detection, it is not of interest to uncover such ability differences as one is only interested in a latent group structure regarding the item difficulties. To avoid any confounding effect of ability differences (or impact), a new score distribution for the Rasch mixture model is introduced here. It ensures the estimation of the Rasch mixture model to be independent of the ability distribution and thus restricts the mixture to be sensitive to latent structure in the item difficulties only. Its usefulness is demonstrated in a simulation study and its application is illustrated in a study of verbal aggression.

Keywords: mixed Rasch model, Rasch mixture model, DIF detection, score distribution.

1. Introduction

Based on the Rasch model (Rasch 1960), Rost (1990) introduced what he called the "mixed Rasch model", a combination of a latent class approach and a latent trait approach to model qualitative and quantitative ability differences. As suggested by Rost (1990), it can also be used to examine the fit of the Rasch model and check for violations of measurement invariance such as differential item functioning (DIF). Since the model assumes latent classes for which separate Rasch models hold, it can be employed to validate a psychological test or questionnaire: if a model with two or more latent classes fits better than a model with one latent class, measurement invariance is violated and a single Rasch model is not suitable because several latent classes are present in the data which require separate Rasch models with separate sets of item difficulties. These classes are latent in the sense that they are not determined by covariates.

As the model assesses a questionnaire – or instrument as it will be referred to in the following – as a whole, it works similar to a global test like the likelihood ratio (LR) test (Andersen 1972; Gustafsson 1980), not an itemwise test like the Mantel-Haenszel test (Holland and Thayer

This is a preprint of an article published in *Educational and Psychological Measurement*, **75**(2), 208–234. doi:10.1177/0013164414536183

1988). Hence, it is the set of item parameters for all items, which is tested for differences between groups rather than each item parameter being tested separately.

The mixed Rasch model – here called Rasch mixture model to avoid confusion with mixed (effects) models and instead highlight its relation to mixture models – has since been extended by Rost and von Davier (1995) to different score distributions and by Rost (1991) and von Davier and Rost (1995) to polytomous responses. The so-called "mixed ordinal Rasch model" is a mixture of partial credit models (PCM, Masters 1982) and includes a mixture of rating scale models (RSM, Andrich 1978) as a special case.

The original dichotomous model as well as its polytomous version have been applied in a variety of fields. Zickar, Gibby, and Robie (2004) use a mixture PCM to detect faking in personality questionnaires, while Hong and Min (2007) identify three types/classes of depressed behavior by applying a mixture RSM to a self-rating depression scale. Another vast field of application are tests in educational measurement. Baghaei and Carstensen (2013) identify different reader types from a reading comprehension test using a Rasch mixture model. Maij-de Meij, Kelderman, and van der Flier (2010) also apply a Rasch mixture model to identify latent groups in a vocabulary test. Cohen and Bolt (2005) use a Rasch mixture model to detect DIF in a mathematics placement test.

Rasch mixture models constitute a legitimate alternative to DIF tests for manifest variables such as the LR test or the recently proposed Rasch trees (Strobl, Kopf, and Zeileis 2014). These methods are usually used to test DIF based on observed covariates, whereas Maij-de Meij *et al.* (2010) show that mixture models are more suitable to detect DIF if the "true source of bias" is a latent grouping variable. The simulation study by Preinerstorfer and Formann (2011) suggests that parameter recovery works reasonably well for Rasch mixture models. While they did not study in detail the influence of DIF effect size or the effect of different ability distributions, they deem such differences relevant for practical concern but leave it to further research to establish just how strongly they influence estimation accuracy.

As the Rasch model is based on two aspects, subject ability and item difficulty, Rasch mixture models are sensitive not only to differences in the item difficulties – as in DIF – but also to differences in abilities. Such differences in abilities are usually called impact and do not infringe on measurement invariance (Ackerman 1992). In practice, when developing a psychological test, one often follows two main steps. First, the item parameters are estimated, e.g., by means of the conditional maximum likelihood (CML) approach, checked for model violations and problematic items are possibly excluded or modified. Second, the final set of items is used to estimate person abilities. The main advantage of the CML approach is that, for a single Rasch model, the estimation and check of item difficulties are (conditionally) independent of the abilities and their distribution. Other global assessment methods like the LR test and the Rasch trees are also based on the CML approach to achieve such independence. However, in a Rasch mixture model, the estimation of the item difficulties is not independent of the ability distribution, even when employing the CML approach. DeMars and Lau (2011) find that a difference in mean ability between DIF groups affects the estimation of the DIF effect sizes. Similarly, other DIF detection methods are also affected by impact, e.g., inflated type I error rates occur in the Mantel-Haenszel and logistic regression procedures if impact is present (Li, Brooks, and Johanson 2012; DeMars 2010).

When using a Rasch mixture model for DIF detection, an influence of impact alone on the mixture is undesirable as the goal is to uncover DIF groups based on item difficulties, not

impact groups based on abilities. To avoid such confounding effects of impact, we propose a new version of the Rasch mixture model specifically designed to detect DIF, which allows for the transfer of the crucial property of CML from a single Rasch model to the mixture: estimation and testing of item difficulties is independent of the abilities and their distribution.

A simulation study is conducted to illustrate how previously suggested versions and this new version of the Rasch mixture model react to impact, either alone or in combination with DIF, and how this affects the suitability of the Rasch mixture model as a DIF detection method.

In the following, we briefly discuss the Rasch model and Rasch mixture models to explain why the latter are sensitive to the specification of the score distribution despite employing a conditional maximum likelihood approach for estimation. This Section 2 is concluded with our suggested new score distribution. We illustrate and discuss the behavior of Rasch mixture models with different options for the score distribution in a Monte Carlo study in Section 3. The suggested approach for DIF detection via Rasch mixture models is illustrated through an empirical application to a study on verbally aggressive behavior in Section 4. Concluding remarks are provided in Section 5.

2. Theory

2.1. The Rasch model

The Rasch model, introduced by Georg Rasch (1960), models the probability for a binary response $y_{ij} \in \{0, 1\}$ by subject i to item j as dependent on the subject's ability θ_i and the item's difficulty β_j. Assuming independence between items given the subject, the probability for observing a vector $y_i = (y_{i1}, \ldots, y_{im})^\top$ with responses to all m items by subject i can be written as

$$P(Y_i = y_i|\theta_i, \beta) \;=\; \prod_{j=1}^{m} \frac{\exp\{y_{ij}(\theta_i - \beta_j)\}}{1 + \exp\{\theta_i - \beta_j\}}, \tag{1}$$

depending on the subject's ability θ_i and the vector of all item difficulties $\beta = (\beta_1, \ldots, \beta_m)^\top$. Capital letters denote random variables and lower case letters denote their realizations.

Since joint maximum likelihood (JML) estimation of all abilities and difficulties is not consistent for a fixed number of items m (Molenaar 1995), conditional maximum likelihood (CML) estimation is employed here. This exploits that the number of correctly scored items, the so-called raw score $R_i = \sum_{j=1}^{m} Y_{ij}$, is a sufficient statistic for the ability θ_i (Molenaar 1995). Therefore, the answer probability from Equation 1 can be split into two factors where the first factor is conditionally independent of θ_i:

$$\begin{aligned} P(Y_i = y_i|\theta_i, \beta) &= P(Y_i = y_i|r_i, \theta_i, \beta)\; P(R_i = r_i|\theta_i, \beta) \\ &= \underbrace{P(Y_i = y_i|r_i, \beta)}_{h(y_i|r_i,\beta)}\; \underbrace{P(R_i = r_i|\theta_i, \beta)}_{g(r_i|\theta_i,\beta)}. \end{aligned}$$

Due to this separation, consistent estimates of the item parameters β can be obtained by maximizing only the conditional part of the likelihood $h(\cdot)$:

$$h(y_i|r_i, \beta) \;=\; \frac{\exp\{-\sum_{j=1}^{m} y_{ij}\beta_j\}}{\gamma_{r_i}(\beta)}, \tag{2}$$

with $\gamma_j(\cdot)$ denoting the elementary symmetric function of order j. The resulting CML estimates $\hat{\beta}$ are consistent, asymptotically normal, and asymptotically efficient (Molenaar 1995).

If not only the conditional likelihood but the full likelihood is of interest – as in Rasch mixture models – then the score distribution $g(\cdot)$ needs to be specified as well. The approach used by Rost (1990) and Rost and von Davier (1995) is to employ some distribution for the raw scores r_i based on a set of auxiliary parameters δ. Then the probability density function for y_i can be written as:

$$f(y_i|\beta, \delta) \;=\; h(y_i|r_i, \beta)\; g(r_i|\delta). \tag{3}$$

Based on this density, the following subsections first introduce mixture Rasch models in general and then discuss several choices for $g(\cdot)$. CML estimation is used throughout for estimating the Rasch model, i.e., the conditional likelihood $h(\cdot)$ is always specified by Equation 2.

2.2. Rasch mixture models

Mixture models are essentially a weighted sum over several components, i.e., here over several Rasch models. Using the Rasch model density function from Equation 3, the likelihood $L(\cdot)$ of a Rasch mixture model with K components for data from n respondents is given by

$$\begin{aligned} L(\pi^{(1)}, \ldots, \pi^{(K)}, \beta^{(1)}, \ldots, \beta^{(K)}, \delta^{(1)}, \ldots, \delta^{(K)}) \;&=\; \prod_{i=1}^{n} \sum_{k=1}^{K} \pi^{(k)} f(y_i|\beta^{(k)}, \delta^{(k)}) \\ &=\; \prod_{i=1}^{n} \sum_{k=1}^{K} \pi^{(k)} h(y_i|r_i, \beta^{(k)})\; g(r_i|\delta^{(k)}). \end{aligned} \tag{4}$$

where the (k)-superscript denotes the component-specific parameters: the component weight $\pi^{(k)}$, the component-specific item parameters $\beta^{(k)}$, and the component-specific score parameters $\delta^{(k)}$ for $k = 1, \ldots, K$.

This kind of likelihood can be maximized via the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) which alternates between maximizing the component-specific likelihoods for obtaining parameter estimates and computing expectations for each observations belonging to each cluster.

More formally, given (initial) estimates for the model parameters $\hat{\pi}^{(k)}$, $\hat{\beta}^{(k)}$, $\hat{\delta}^{(k)}$ for all components $k = 1, \ldots, K$, posterior probabilities of each observation i belonging to a component, or latent class, k are calculated in the E-step. This is simply i's relative contribution to component k compared to the sum of all its contributions:

$$\hat{p}_{ik} \;=\; \frac{\hat{\pi}^{(k)}\; f(y_i|\hat{\beta}^{(k)}, \hat{\delta}^{(k)})}{\sum_{\ell=1}^{K} \hat{\pi}^{(\ell)}\; f(y_i|\hat{\beta}^{(\ell)}, \hat{\delta}^{(\ell)})} \;=\; \frac{\hat{\pi}^{(k)}\; h(y_i|r_i, \hat{\beta}^{(k)})\; g(r_i|\hat{\delta}^{(k)})}{\sum_{\ell=1}^{K} \hat{\pi}^{(\ell)}\; h(y_i|r_i, \hat{\beta}^{(\ell)})\; g(r_i|\hat{\delta}^{(\ell)})}\,. \tag{5}$$

In the M-step of the algorithm, these posterior probabilities are used as the weights in a weighted ML estimation of the model parameters. This way, an observation deemed unlikely to belong to a certain latent class does not contribute strongly to its estimation. Estimation can be done separately for each latent class. Using CML estimation for the Rasch Model, the estimation of item and score parameters can again be done separately. For all components

$k = 1, \ldots, K$:

$$\begin{aligned}(\hat{\beta}^{(k)}, \hat{\delta}^{(k)}) &= \underset{\beta^{(k)}, \delta^{(k)}}{\operatorname{argmax}} \sum_{i=1}^{n} \hat{p}_{ik} \log f(y_i | \beta^{(k)}, \delta^{(k)}) \\ &= \left\{ \underset{\beta^{(k)}}{\operatorname{argmax}} \sum_{i=1}^{n} \hat{p}_{ik} \log h(y_i | r_i, \beta^{(k)}); \;\; \underset{\delta^{(k)}}{\operatorname{argmax}} \sum_{i=1}^{n} \hat{p}_{ik} \log g(r_i | \delta^{(k)}) \right\}. \end{aligned} \tag{6}$$

Estimates of the class probabilities can be obtained from the posterior probabilities by averaging:

$$\hat{\pi}^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{ik}. \tag{7}$$

The E-step (Equation 5) and M-step (Equations 6 and 7) are iterated until convergence, always updating either the weights based on current estimates for the model parameters or vice versa.

Note that the above implicitly assumes that the number of latent classes K is given or known. However, this is typically not the case in practice and K needs to be chosen based on the data. As K is not a model parameter – regularity conditions for the likelihood ratio test are not fulfilled (McLachlan and Peel 2000, Chapter 6.4) – it is often chosen via some information criterion that balances goodness of fit (via the likelihood) with a penalty for the number of model parameters. Since the various information criteria differ in their penalty term, the decision which model is considered "best" may depend on the information criterion chosen. In the following, the BIC (Bayesian information criterion, Schwarz 1978) is used, which Li, Cohen, Kim, and Cho (2009) found to be a suitable model selection method for dichotomous mixture item response theory models. Note that this is not a formal significance test because one does not control a type I error rate.

2.3. Score distribution

In a single Rasch model, the estimation of the item parameters is invariant to the score distribution because of the separation in Equation 3. In the mixture context, this invariance property holds only *given the weights* in Equation 6. However, these posterior weights depend on the full Rasch likelihood, including the score distribution (Equation 5). Therefore, the estimation of the item parameters in a Rasch mixture model is *not* independent of the score distribution for $K > 1$, even if the CML approach is employed. Hence, it is important to consider the specification of the score distribution when estimating Rasch mixture models and to assess the consequences of potential misspecifications.

Saturated and mean-variance specification

In his introduction of the Rasch mixture model, Rost (1990) suggests a discrete probability distribution on the scores with a separate parameter for each possible score. This requires $m - 2$ parameters per latent class as the probabilities need to sum to 1 (and the extreme scores, $r = 0$ and $r = m$, do not contribute to the likelihood).

Realizing that this saturated specification requires a potentially rather large number of parameters, Rost and von Davier (1995) suggest a parametric distribution with one parameter each for mean and variance.

Details on both specifications can be found in Rost (1990) and Rost and von Davier (1995), respectively. Here, the notation of Frick, Strobl, Leisch, and Zeileis (2012) is adopted, which expresses both specifications in a unified way through a conditional logit model for the score $r = 1, \ldots, m-1$:

$$g(r|\delta^{(k)}) \;=\; \frac{\exp\{z_r^\top \delta^{(k)}\}}{\sum_{j=1}^{m-1} \exp\{z_j^\top \delta^{(k)}\}},$$

with different choices for z_r leading to the saturated and mean-variance specification, respectively. For the former, the regressor vector is $(m-2)$-dimensional with

$$z_r = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$$

and the 1 at position $r-1$. Consequently, if $r = 1$, z_r is a vector of zeros. For the mean-variance specification, the regressor vector is 2-dimensional and given by

$$z_r = \left(\frac{r}{m}, \frac{4r(m-r)}{m^2}\right)^\top.$$

Restricted specification

In the following we suggest a new specification of the score distribution in the Rasch mixture model, which aims at obtaining independence of the item parameter estimates from the specification of the score distribution and therefore enabling the Rasch mixture model to distinguish between DIF and impact. Other global DIF detection methods like the LR test and Rasch trees are able to make this distinction (Ankenmann, Witt, and Dunbar 1999; Strobl *et al.* 2014) because they are based only on the conditional part of the likelihood (Equation 2). Analogously, we suggest a mixture of only this conditional part rather than the full likelihood (Equation 3) of the Rasch model so that the mixture model will only be influenced by differences in the item parameters.

Mixing only the conditional likelihood $h(\cdot)$ means that the sum over the K latent classes in the likelihood of the Rasch mixture model in Equation 4 only applies to $h(\cdot)$ but not to the score distribution $g(\cdot)$. The mixture is then only based on latent structure in the item difficulties, not on latent structure in both difficulties and scores. Moreover, such a Rasch mixture model based only on the conditional likelihood without any score distribution is equivalent to a Rasch mixture model where the score distribution is independent of the latent class $k = 1, \ldots, K$:

$$g(r|\delta^{(k)}) = g(r|\delta) \qquad (k = 1, \ldots, K),$$

because then the factor $g(r|\delta)$ is a constant that can be moved out of the sum over the components k in Equation 4. Consequently, compared to the case without any score distribution, the log-likelihood just changes by an additional constant without component-specific parameters. In either case, the estimation of the component-specific parameters item parameters as well as the selection of the number of components K is independent of the specification of the score distribution.

This equivalence and independence from the score distribution can also be seen easily from the definition of the posterior weights (Equation 5): If restricted, $g(\cdot)$ can be moved out of

the sum and then cancels out, preserving only the dependence on $h(\cdot)$. Thus, the $\hat{p}_{ik}$ depend only on $\hat{\pi}^{(k)}$ and $\hat{\beta}^{(k)}$ but not $\hat{\delta}^{(k)}$. Therefore, the component weights and component-specific item parameters can be estimated without any specification of the score distribution.

Subsequently, we adopt the restricted perspective rather than omitting $g(\cdot)$ completely, when we want to obtain a mixture model where the mixture is independent of the score distribution. From a statistical point of view this facilitates comparisons of the restricted Rasch mixture model with the corresponding unrestricted counterpart.

Overview

The different specifications of the score distribution vary in their properties and implications for the whole Rasch mixture model.

- The saturated model is very flexible. It can model any shape and is thus never misspecified. However, it needs a potentially large number of parameters which can be challenging in model estimation and selection.
- The mean-variance specification of the score model is more parsimonious as it only requires two parameters per latent class. While this is convenient for model fit and selection, it also comes at a cost: since it can only model unimodal or U-shaped distributions (see Rost and von Davier 1995), it is partially misspecified if the score distribution is actually multimodal.
- A restricted score model is even more parsimonious. Therefore, the same advantages in model fit and selection apply. Furthermore, it is invariant to the latent structure in the score distribution. If a Rasch mixture model is used for DIF detection, this is favorable as only differences in the item difficulties influence the mixture. However, it is partially misspecified if the latent structure in the scores and item difficulties coincides.

3. Monte Carlo study

The simple question *"DIF or no DIF?"* leads to the question whether the Rasch mixture model is suitable as a tool to detect such violations of measurement invariance.

As the score distribution influences the estimation of the Rasch mixture model in general, it is of particular interest how it influences the estimation of the number of latent classes, the measure used to determine Rasch scalability.

3.1. Motivational example

As a motivation for the simulation design, consider the following example: The instrument is a knowledge test which is administered to students from two different types of schools and who have been prepared by one of two different courses for the knowledge test. Either of the two groupings might be the source of DIF (or impact). If the groupings are available as covariates to the item responses of the students, then a test for DIF between either school types or course types can be easily carried out using the LR test. However, if the groupings are not available (or even observed) as covariates, then a DIF test is still possible by means

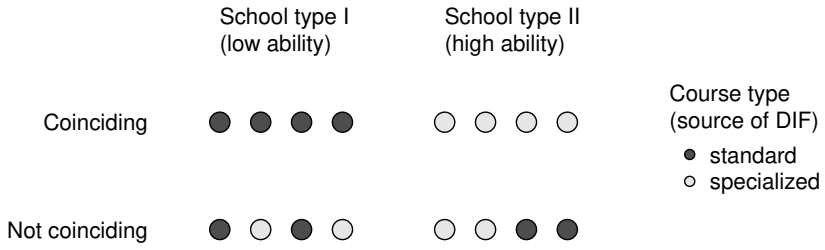

Figure 1: Grouping structure in the motivational example.

of the Rasch mixture model. The performance of such a DIF assessment is investigated in our simulation study for different effects of school and course type, respectively.

In the following we assume that the school type is linked to ability difference (i.e., impact but not DIF) while the course type is the source of DIF (but not impact). This can be motivated in the following way (see also Figure 1): When the students from the two school types differ in their mean ability, this is impact between these two groups. The courses might be a standard course and a new specialized course. While the standard course covers all topics of the test equally, the specialized course gives more emphasis to a relatively advanced topic and due to time constraints less emphasis to a relatively basic topic. This may lead to DIF between the students in the standard and the specialized course. See the left panel of Figure 2 for illustrative item profiles of the standard course (in dark gray) and the specialized course (in light gray).

Finally, the ability groups by school and the DIF groups by course can either coincide or not. If all students in the first school type are being taught the standard course while all students in the second school type are being taught the specialized course, the DIF groups *coincide* with the ability groups. The DIF and ability groups do *not coincide* but only overlap partly if both course types are taught in both school types: each DIF group (based on the type of course taught) consists of a mix of students from both schools and therefore from both ability groups. An illustration of coinciding and not coinciding ability and DIF groups is provided in the upper and lower row of Figure 1, respectively. Ability groups, based on school type, are shown in the columns, while DIF groups, based on course type, are illustrated with dark and light gray for the standard course and specialized course, respectively. This difference of coinciding or not coinciding DIF and ability groups might have an influence on the Rasch mixture model's ability to detect the DIF because in the former case the score distributions differ between the two DIF groups while in the latter case they do not.

Subsequently, a Monte Carlo study is carried out to investigate how the Rasch mixture model performs in situations where such groupings are present in the underlying data-generating process but are not available as observed covariates. Moreover, we vary whether or not all students come from the same school type (i.e., from the same ability distribution), whether or not all students receive the standard course (i.e., whether there is DIF), and whether both school types use the same or different courses (i.e., whether the groupings coincide or not). For all computations, the R system for statistical computing (R Core Team 2013) is used along with the add-on packages **psychomix** (Frick *et al.* 2012) and **clv** (Nieweglowski 2013).

Scenario		Latent class I		Latent class II	
		Mean abilities	Difficulties	Mean abilities	Difficulties
No impact ($\Theta = 0$)					
1	no DIF ($\Delta = 0$)	$\{0\}$	β^I	—	—
2	DIF ($\Delta > 0$)	$\{0\}$	β^I	$\{0\}$	β^{II}
Impact ($\Theta > 0$)					
3	no DIF ($\Delta = 0$)	$\{-\Theta/2, +\Theta/2\}$	β^I	—	—
4	DIF ($\Delta > 0$), not coinciding	$\{-\Theta/2, +\Theta/2\}$	β^I	$\{-\Theta/2, +\Theta/2\}$	β^{II}
5	DIF ($\Delta > 0$), coinciding	$\{-\Theta/2\}$	β^I	$\{+\Theta/2\}$	β^{II}

Table 1: Simulation design. The latent-class-specific item parameters β^I and β^{II} differ by Δ for two elements and thus coincide for $\Delta = 0$, leaving only a single latent class.

3.2. Simulation design

The simulation design combines ideas from the motivational example with aspects from the simulation study conducted by Rost (1990). Similar to the original simulation study, the item parameters represent an instrument with increasingly difficult items. Here, 20 items are employed with corresponding item parameters β^I which follow a sequence from -1.9 to 1.9 with increments of 0.2 and hence sum to zero.

$$\begin{aligned} \beta^I &= (-1.9, -1.7, \ldots, 1.7, 1.9)^\top \\ \beta^{II} &= (-1.9, -1.7, \ldots, -1.1 + \Delta, \ldots, 1.1 - \Delta, \ldots, 1.7, 1.9)^\top \end{aligned}$$

To introduce DIF, a second set of item parameters β^{II} is considered where items 5 and 16 are changed by $\pm\Delta$. This approach is similar in spirit to that of Rost (1990) – who reverses the full sequence of item parameters to generate DIF – but allows for gradually changing from small to large DIF effect sizes. Subject abilities are drawn with equal weights from two normal distributions with means $-\Theta/2$ and $+\Theta/2$ and standard deviation 0.3, thus creating a sample with two groups of subjects: one group with a lower mean ability and one with a higher mean ability.

In the simulations below, the DIF effect size Δ ranges from 0 to 4 in steps of 0.2

$$\Delta \in \{0, 0.2, \ldots, 4\}$$

while the impact Θ covers the same range in steps of 0.4:

$$\Theta \in \{0, 0.4, \ldots, 4\}.$$

Impact and DIF, or lack thereof, can be combined in several ways. Table 1 provides an overview and Figures 2, 3, and 4 show illustrations. In the following, the different combinations of impact and DIF are explained in more detail and connected to the motivational example:

- If the simulation parameter Δ for the DIF effect size is set to zero, both sets of item parameters, β^I and β^{II}, are identical and no DIF is present. Since CML is employed, model selection and parameter estimation is typically expected to be independent of whether or not an impact is present (Scenario 1 and 3 in Table 1).

  In the example: Only the standard course is taught and hence no DIF exists.

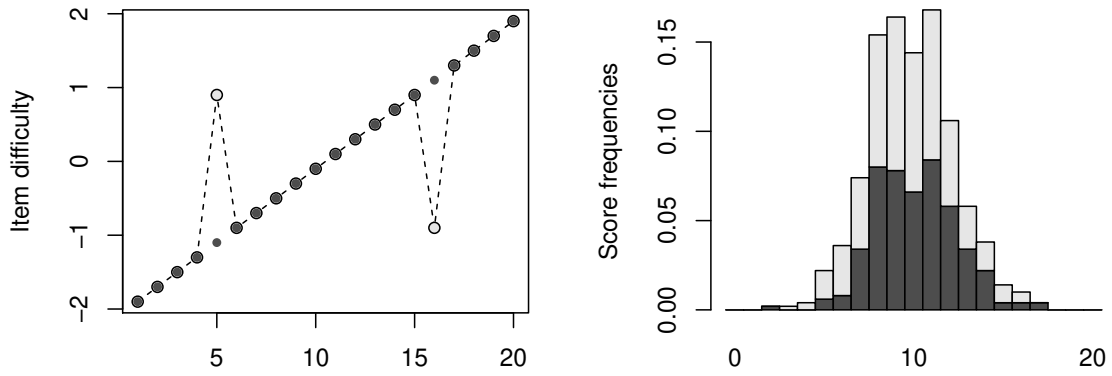

Figure 2: Scenario 2. Left: Item difficulties with DIF ($\Delta = 2$). Right: Stacked histogram of unimodal score distribution with homogeneous abilities ($\Theta = 0$).

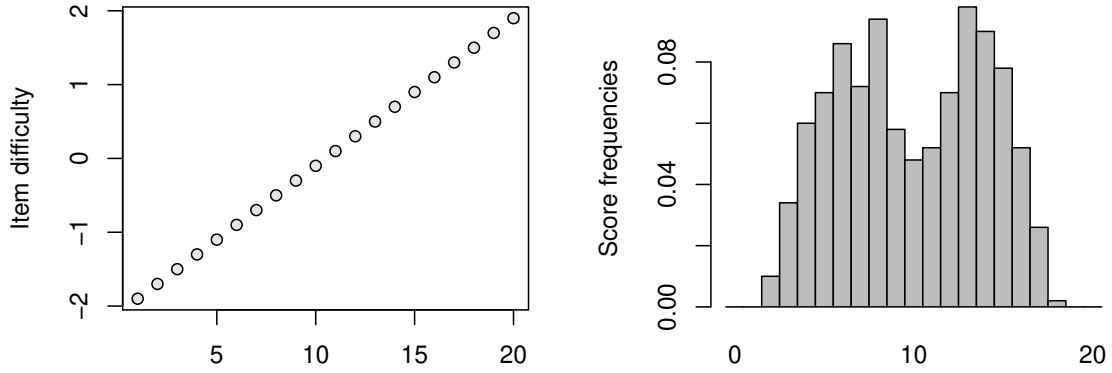

Figure 3: Scenario 3. Left: Item difficulties without DIF ($\Delta = 0$). Right: Histogram of bimodal score distribution with impact ($\Theta = 2$).

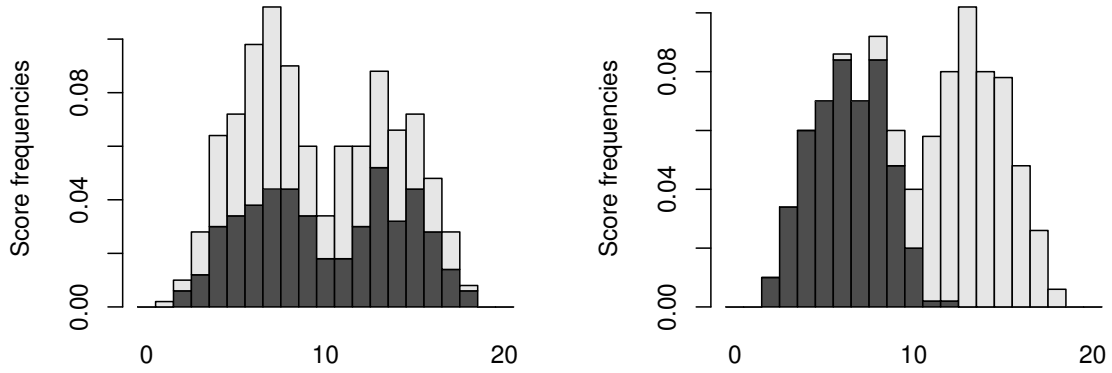

Figure 4: Stacked histograms of score distributions for Scenarios 4 (left) and 5 (right) with DIF ($\Delta = 2$). Left: impact and DIF, not coinciding ($\Theta = 2$). Right: impact and DIF, coinciding ($\Theta = 2$). For item difficulties see Figure 2 (left).

- If $\Delta > 0$, the item parameter set β^{II} is different from β^{I}. Hence, there is DIF and two latent classes exist (Scenarios 2, 4, and 5). Both classes are chosen to be of equal size in this case. For an illustration see the left panel of Figure 2.

  In the example: Both courses are taught, thus leading to DIF. The standard course corresponds to the straight line as the item profile while the specialized course corresponds to the spiked item profile with relatively difficult item 16 being easier and the relatively easy item 5 being more difficult for students in this specialized course than for students in the standard course.

- If the simulation parameter Θ for the impact is set to zero (Scenarios 1 and 2), then the resulting score distribution is unimodal. For an illustration of such a unimodal score distribution see the right panel of Figure 2. This histogram illustrates specifically Scenario 2 where no impact is present but DIF exists. The histogram is shaded in light and dark gray for the two DIF groups and thus to be read like a "stacked histogram".

  In the example: All students are from the same school and hence there is no impact. However, both types of courses may be taught in this one school, thus leading to DIF as in Scenario 2.

- If $\Theta > 0$, subject abilities are sampled with equal weights from two normal distributions with means $\{-\Theta/2, +\Theta/2\}$, thus generating impact. When no DIF is included (Scenario 3), the resulting score distribution moves from being unimodal to being bimodal with increasing Θ. The two modi of high and low scores represent the two groups of subjects with high and low mean abilities, respectively. However, only a medium gray is used to shade the illustrating histogram in Figure 3 as *no DIF groups* are present.

  In the example: Only the standard course is taught in both school types. Hence no DIF is present but impact between the school types.

- If there is DIF (i.e., $\Delta > 0$) in addition to impact (i.e., $\Theta > 0$), subjects can be grouped both according to mean ability (high vs. low) and difficulty (straight vs. spiked profile in β^{I} and β^{II}, respectively).

  These groups can *coincide*: For subjects with low mean ability $-\Theta/2$, item difficulties β^{I} hold, while for subjects with high mean ability $+\Theta/2$, item difficulties β^{II} hold. This is simulated in Scenario 5 and labeled *Impact and DIF, coinciding*. The resulting score distribution is illustrated in the right panel of Figure 4. Subjects for whom item difficulties β^{I} hold are shaded in dark gray and as they also have lower mean abilities, their scores are all relatively low. Conversely, subjects for whom item difficulties β^{II} hold are shaded in light gray and as they also have higher mean abilities, their scores are all relatively high.

  Additionally, the DIF groups and ability groups can also *not coincide*: Subjects in either DIF group may stem from both ability groups, not just one. This is simulated in Scenario 4 and labeled *Impact and DIF, not coinciding*. The resulting score distribution is illustrated in the left panel of Figure 4. Again, subjects for whom item difficulties β^{I} and β^{II} hold are shaded in dark and light gray, respectively. As subjects stem from both ability groups (high vs. low abilities), both score distributions are bimodal.

  In the example: Students from both school types and from both course types are considered, thus leading to both impact and DIF. Either both courses are taught at both

schools (Scenario 4, not coinciding) or the standard course is only taught in the first school and the specialized course is only taught at the second school (Scenario 5, coinciding).

Note that Scenario 1 is a special case of Scenario 2 where Δ is reduced to zero as well as a special case of Scenario 3 where Θ is reduced to zero. Therefore, in the following, Scenario 1 is not inspected separately but included in both the setting of *No impact with DIF* (Scenario 2) and the setting of *Impact without DIF* (Scenario 3) as a reference point. Similarly, Scenarios 4 and 5 both can be reduced to Scenario 3 if Δ is set to zero. It is therefore also included in both the setting of *Impact and DIF, not coinciding* (Scenario 4) and the setting of *Impact and DIF, coinciding* (Scenario 5) as a reference point.

For each considered combination of Δ and Θ, 500 datasets of 500 observations each are generated. Larger numbers of datasets or observations lead to very similar results. Observations with raw scores of 0 or m are removed from the dataset as they do not contribute to the estimation of the Rasch mixture model (Rost 1990). For each dataset, Rasch mixture models for each of the saturated, mean-variance, and restricted score specifications are fitted for $K = 1, 2, 3$.

3.3. False alarm rate and hit rate

The main objective here is to determine how suitable a Rasch mixture model, with various choices for the score model, is to recognize DIF or the lack thereof.

For each dataset and type of score model, models with $K = 1, 2, 3$ latent classes are fitted and the $\hat{K}$ associated with the minimum BIC is selected. Choosing one latent class ($\hat{K} = 1$) then corresponds to assuming measurement invariance while choosing more than one latent class ($\hat{K} > 1$) corresponds to assuming violations of measurement invariance. While Rasch mixture models do not constitute a formal significance test, the empirical proportion among the 500 datasets with $\hat{K} > 1$ corresponds in essence to the power of DIF detection if $\Delta > 0$ (and thus two true latent classes exist) and to the associated type I error of a corresponding test if $\Delta = 0$ (and thus only one true latent class exists). If the rate corresponds to power, it will be referred to as *hit rate* whereas if it corresponds to a type I error it will be referred to as *false alarm rate*.

In the following subsections, the key results of the simulation study will be visualized. The exact rates for all conditions are included as a dataset in the R package **psychomix**, for details see the section on computational details.

Scenario 2: No impact with DIF

This scenario is investigated as a case of DIF that should be fairly simple to detect. There is no impact as abilities are homogeneous across all subjects so the only latent structure to detect is the group membership based on the two item profiles. This latent structure is made increasingly easy to detect by increasing the difference between the item difficulties for both latent groups. In the graphical representation of the item parameters (left panel of Figure 2) this corresponds to enlarging the spikes in the item profile.

Figure 5 shows how the rate of choosing a model with more than one latent class ($\hat{K} > 1$) increases along with the DIF effect size Δ. At $\Delta = 0$, this is a false alarm rate. It is around 7% for the saturated model and very close to zero for the mean-variance and the saturated

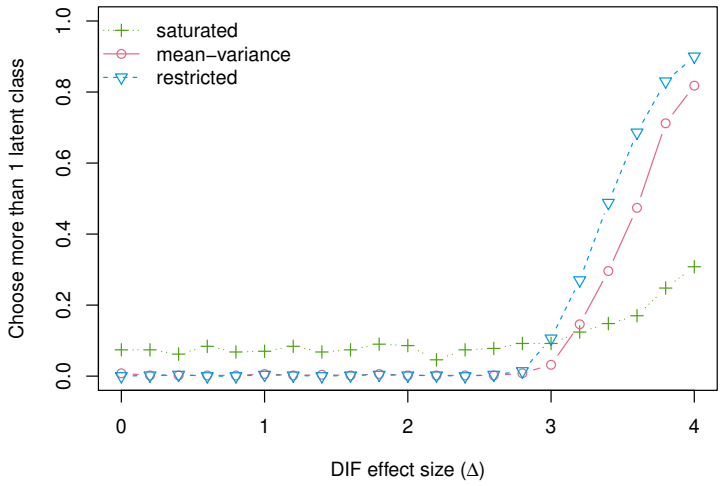

Figure 5: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 2 (DIF without impact, i.e., $\Theta = 0$).

score model ($< 1\%$). With increasing $\Delta > 0$, the rate is a hit rate. For low values of Δ the two more parsimonious versions of the Rasch mixture model (with mean-variance and restricted score distribution) are not able to pick up the DIF but at around $\Delta = 3$ the hit rate for the two models increases and almost approaches 1 at $\Delta = 4$. Not surprisingly, the restricted score specification performs somewhat better because in fact the raw score distributions do not differ between the two latent classes. The baseline hit rate of the saturated model for low values of Δ is the same as the false alarm rate for $\Delta = 0$. It only increases beyond the same threshold ($\Delta = 3$) as the hit rate of the other two models. However, its rate is much lower compared to the other two score model (only around 30%). The reason is that it requires 18 additional score parameters for an additional latent class which is "too costly" in terms of BIC. Hence, $\hat{K} = 1$ is chosen for most Rasch mixture models using a saturated score distribution.

The number of iterations in the EM algorithm which are necessary for the estimation to converge is much lower for the mean-variance and the restricted model than for the saturated model. Since the estimation of the saturated model is more extensive due to the higher number of parameters required by this model, it does not converge in about 10% of the cases before reaching the maximum number of iterations which was set to 400. The mean-variance and saturated model usually converge within the first 200 iterations.

Brief summary: The mean-variance and restricted model have higher hit rates than the saturated model in the absence of impact.

Scenario 3: Impact without DIF

Preferably, a Rasch mixture model should not only detect latent classes if the assumption of measurement invariance is violated but it should also indicate a lack of latent structure if indeed the assumption holds. In this scenario, the subjects all stem from the same class, meaning each item is of the same difficulty for every subject. However, subject abilities are simulated with impact resulting in a bimodal score distribution as illustrated in Figure 3.

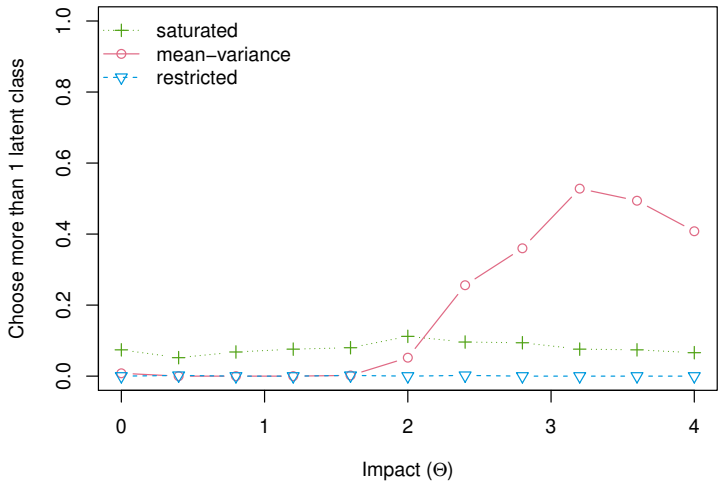

Figure 6: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 3 (impact without DIF, i.e., $\Delta = 0$).

Here, the rate of choosing more than one latent class can be interpreted as a false alarm rate (Figure 6). The restricted score model is invariant against any latent structure in the score distribution and thus almost always ($\leq 0.2\%$) suggests $\hat{K} = 1$ latent class based on the DIF-free item difficulties. The rate does not approach any specific significance level as the Rasch mixture model, regardless of the employed score distribution, is not a formal significance test. The saturated model also picks $\hat{K} = 1$ in most of the simulation. This might be due to its general reluctance to choose more than one latent class as illustrated in Figure 5 or the circumstance that it can assume any shape (including bimodal patterns). However, the mean-variance score distribution can only model unimodal or U-shaped distributions as mentioned above. Hence, with increasing impact and thus increasingly well-separated modes in the score distribution, the Rasch mixture model with this score specification suggests $\hat{K} > 1$ latent classes in up to 53% of the cases. Note, however, that these latent classes do not represent the DIF groups (as there are none) but rather groups of subjects with high vs. low abilities. While this may be acceptable (albeit unnecessarily complex) from a statistical mixture modeling perspective, it is misleading from a psychometric point of view if the aim is DIF detection. Only one Rasch model needs to be estimated for this type of data, consistent item parameter estimates can be obtained via CML and all observations can be scaled in the same way.

Brief summary: If measurement invariance holds but ability differences are present, the mean-variance model exhibits a high false alarm rate while the saturated and restricted model are not affected.

Scenario 4: Impact and DIF, not coinciding

In this scenario, there is DIF (and thus two true latent classes) if $\Delta > 0$. Again, Scenario 3 with $\Delta = 0$ (and thus without DIF) is included as a reference point. However, unlike in Scenario 2, the abilities within the latent classes are not homogeneous but two ability groups exist, which do not coincide with the two DIF groups. Nonetheless, the score distribution is

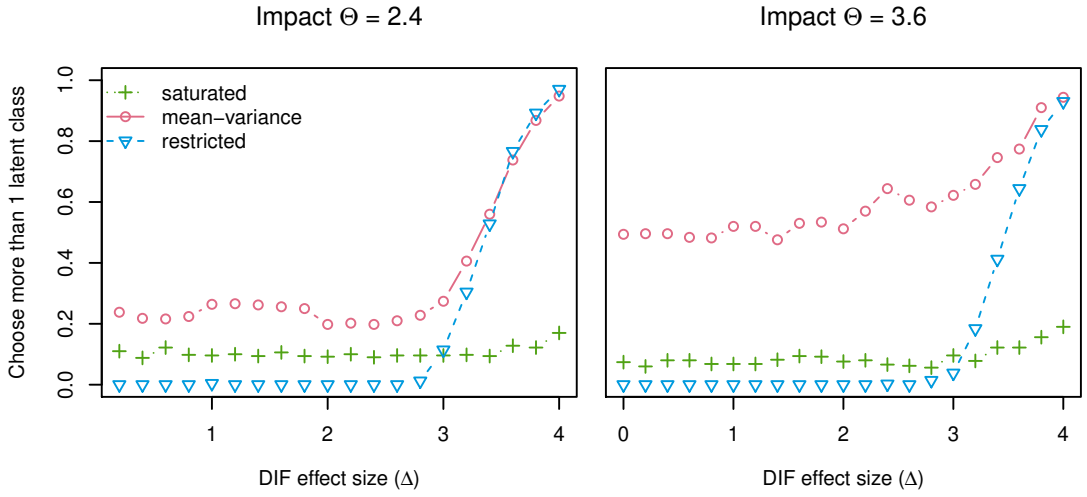

Figure 7: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 4 (impact and DIF, not coinciding).

the same across both latent classes (illustrated in the left panel of Figure 4).

Figure 7 again shows the rate of choosing $\hat{K} > 1$ for increasing DIF effect size Δ for two levels of impact ($\Theta = 2.4$ and 3.6), exemplary for medium and high impact. If impact is small (e.g., $\Theta = 0.4$), the rates are very similar to the case of completely homogeneous abilities without impact (Figure 5 with $\Theta = 0$) and thus not visualized here. While the rates for the restricted and the saturated score model do not change substantially for an increased impact ($\Theta = 2.4$ and 3.6), the mean-variance model is influenced by this change in ability differences. While the hit rate is increased to around 20% over the whole range of Δ, the false alarm rate at $\Delta = 0$ is increased to the same extent. Moreover, the hit rate only increases noticeably beyond the initial false alarm rate at around $\Delta = 3$, i.e., the same DIF effect size at which the restricted and mean-variance specifications have an increasing hit rate given homogeneous abilities without impact. Thus, given rather high impact ($\Theta = 3.6$) the hit rate is not driven by the DIF detection but rather the model's tendency to assign subjects with high vs. low abilities into different groups (as already seen in Figure 6).

As Rasch mixture models with $K = 1, 2, 3$ classes are considered, selecting $\hat{K} > 1$ classes can either mean selecting the correct number of $K = 2$ or overselecting $\hat{K} = 3$ classes. For the saturated and restricted specifications overselection is rare (occurring with rates of less than 9% or less than 1%, respectively). However, similar to Scenario 3 overselection is not rare for the mean-variance specification. Figure 8 depicts the rates of selecting $\hat{K} = 2$ and $\hat{K} = 3$ classes, respectively, for increasing Δ at $\Theta = 3.6$. If the chances of finding the correct number of classes increase with the DIF effect size Δ, the rate for overselection ($\hat{K} = 3$) should drop with increasing Δ. For this Scenario 4, denoted with hollow symbols, this rate stays largely the same (around 25%) and even slightly increases beyond this level, starting from around $\Delta = 3$. This illustrates again the pronounced tendency of the mean-variance model for overselection in cases of high impact.

Brief summary: If impact is simulated within DIF groups, the mean-variance model has higher

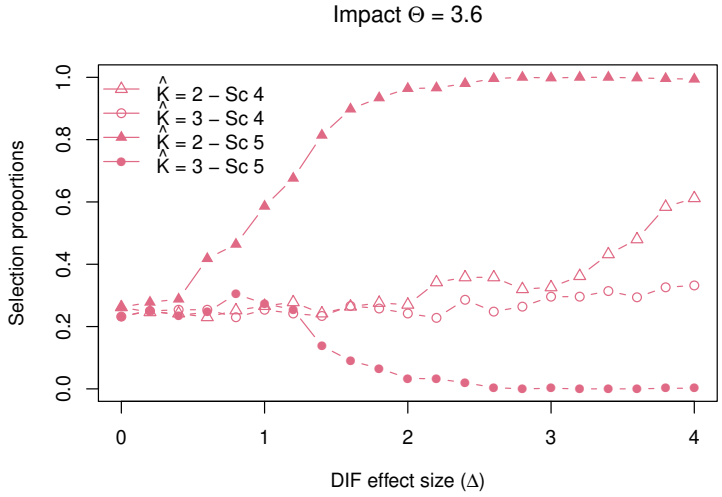

Figure 8: Rates of choosing the correct number of classes ($\hat{K} = 2$) or overselecting the number of classes ($\hat{K} = 3$) for the Rasch mixture model with mean-variance score specification in Scenarios 4 (hollow, impact within DIF groups) and 5 (solid, impact between DIF groups).

hit rates than the saturated and restricted models. However, the latent classes estimated by the mean-variance model are mostly based on ability differences if the DIF effect size is low. If the DIF effect size is high, the mean-variance model tends to overestimate the number of classes.

Scenario 5: Impact and DIF, coinciding

In Scenario 5, there is also DIF (i.e., $\Delta > 0$) and impact. However, in contrast to Scenario 4 the ability and DIF groups coincide (see the right panel of Figure 4). Furthermore, Scenario 3 is included also here as the reference point without DIF ($\Delta = 0$).

Again, small ability differences do not strongly influence the rate of choosing more than one latent class (rates for low levels of impact, such as $\Theta = 0.4$, are similar to those for $\Theta = 0$ as depicted in Figure 5). Recall, both mean-variance and restricted specification have comparable hit rates for DIF detection starting from around $\Delta = 3$ while the saturated specification has lower hit rates.

As impact increases (Figure 9), the hit rates of all models increases as well because the ability differences contain information about the DIF groups: separating subjects with low and high abilities also separates the two DIF groups (not separating subjects within each DIF group as in the previous setting). However, for the mean-variance model these increased hit rates are again coupled with a highly increased false alarm rate at $\Delta = 0$ of 26% and 50% for $\Theta = 2.4$ and 3.6, respectively. The restricted score model, on the other hand, is invariant to latent structure in the score distribution and thus performs similarly as in previous DIF scenarios, suggesting more than one latent class past a certain threshold of DIF intensity, albeit this threshold being a bit lower than when ability groups and DIF groups do not coincide (around $\Delta = 2$). The saturated model detects more than one latent class at a similar rate to the restricted score model for medium or high impact but its estimation converges more slowly and requires more iterations of the EM algorithm than the other two score models.

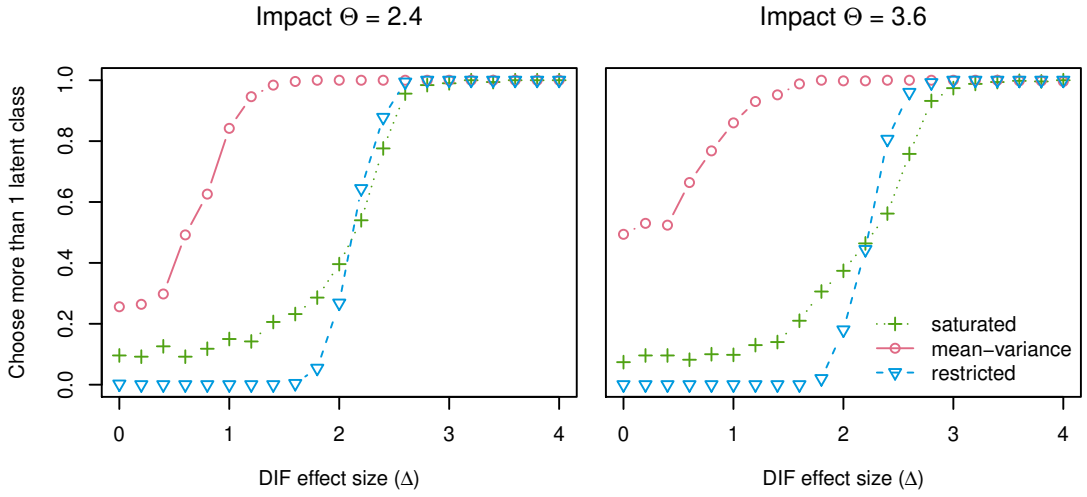

Figure 9: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 5 (impact and DIF, coinciding).

Finally, the potential issue of overselection can be considered again. Figure 8 (solid symbols) shows that this problem disappears for the mean-variance specification if both DIF effect size Δ and impact are large *and* coincide. For the restricted model overselection is again very rare throughout (occurring in less than 1% of all cases) while the saturated model overselects in up to 29% of the datasets.

Brief summary: If abilities differ between DIF groups, the mean-variance model detects the violation of measurement invariance for smaller DIF effect sizes than the saturated and restricted model. While the mean-variance model does not overselect the number of components in this scenario, the high hit rates are connected to a high false alarm rate when no DIF is present but impact is high. This does not affect the other two score models.

3.4. Quality of estimation

Although here the Rasch mixture model is primarily used analogously to a global DIF test, model assessment goes beyond the question whether or not the correct number of latent classes is found. Once the number of latent classes is established/estimated, it is of interest how well the estimated model fits the data. Which groups are found? How well are the parameters estimated? In the context of Rasch mixture models with different score distributions, both of these aspects depend heavily on the posterior probabilities $\hat{p}_{ik}$ (Equation 5) as the estimation of the item parameters depends on the score distribution only through these. If the $\hat{p}_{ik}$ were the same for all three score specifications, the estimated item difficulties were the same as well. Hence, the focus here is on how close the estimated posterior probabilities are to the true latent classes in the data. If the similarity between these is high, CML estimation of the item parameters within the classes will also yield better results for all score models.

This is a standard task in the field of cluster analysis and we adopt the widely used Rand index (Rand 1971) here: Each observation is assigned to the latent class for which its posterior probability is highest yielding an estimated classification of the data which is compared to the

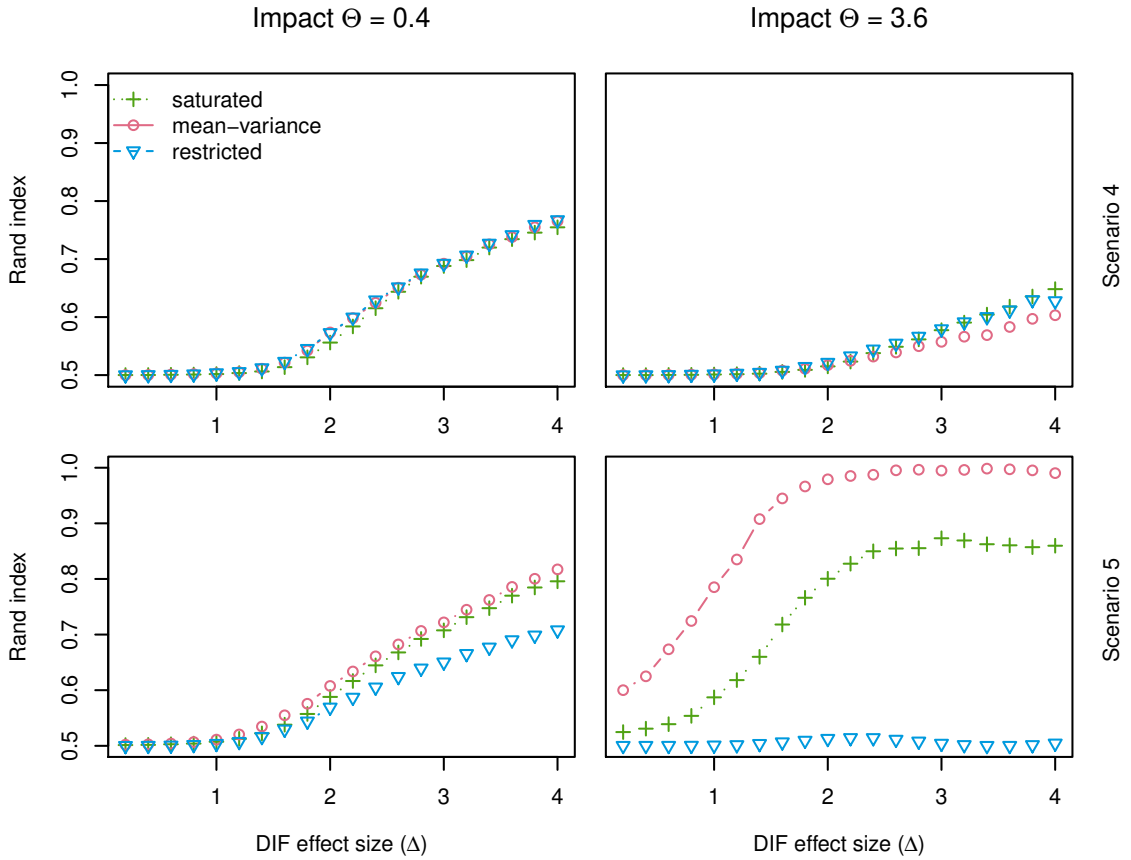

Figure 10: Average Rand index for models with $K = 2$ latent classes. Top row: Scenario 4 (impact and DIF, not coinciding). Bottom row: Scenario 5 (impact and DIF, coinciding).

true classification. For this comparison, pairs of observations are considered. Each pair can either be in the same class in both the true and the estimated classification, in different classes for both classifications or it can be in the same class for one but not the other classification. The Rand index is the proportion of pairs for which both classifications agree. Thus, it can assume values between 0 and 1, indicating total dissimilarity and similarity, respectively.

In the following, the Rand index for models with the true number of $K = 2$ latent classes in Scenarios 4 and 5 (with DIF) is considered. Thus, the question of DIF detection (or model selection) is not investigated again but only the quality of latent class recovery (assuming the number of classes K to be known or correctly selected). The top row of Figure 10 depicts the average Rand index for data from Scenario 4 (impact and DIF, not coinciding). Here, all three score specifications find similarly well matching classifications, while the Rand index generally decreases with increasing impact (left to right panel). In particular, while the mean-variance score model has problems finding the *correct number* of latent classes in this scenario, it only performs slightly worse than the other two specifications in determining the *correct classes* if the number were known. Similarly, if it is provided with the correct number of classes, the saturated model also identifies the correct classes equally well compared to the other models

– despite its difficulties with convergence for higher DIF effect sizes.

However, in Scenario 5 where the score distribution contains information about the DIF groups, the three score specifications perform very differently as the bottom row of Figure 10 shows. Given the correct number of classes, the mean-variance model is most suitable to uncover the true latent classes, yielding Rand indices close to 1 if both DIF effect size and impact are large. The saturated specification follows a similar pattern albeit with poorer results, reaching values of up to 0.87. However, the classifications obtained from the restricted score specification do not match the true groups well in this scenario, remaining below 0.52 if impact is high. The reason is that the restricted score model is partially misspecified as the score distributions differ substantially across DIF groups.

3.5. Summary and implications for practical use

Given various combinations of DIF and ability impact, the score models are differently suitable for the two tasks discussed here – DIF detection and estimation of item parameters in subgroups. Starting with a summary of the results for DIF detection:

- The saturated score model has much lower hit rates than the other two specifications, i.e., violation of measurement invariance remains too often undetected. Only if high impact and high DIF effect sizes coincide does the saturated model perform similarly well as the restricted model.
- The mean-variance model has much higher hit rates. However, if impact is present in the abilities, this specification has highly inflated false alarm rates. Hence, if the mean-variance model selects more than one latent class it is unclear whether this is due to DIF or just varying subject abilities. Thus, measurement invariance might still hold even if more than one latent class is detected.
- The restricted score model also has high hit rates, comparable to the mean-variance model if abilities are rather homogeneous. But unlike the mean-variance specification, its false alarm rate is not distorted by impact. Its performance is not influenced by the ability distribution and detecting more than one latent class reliably indicates DIF, i.e., a violation of measurement invariance.

Hence, if the Rasch mixture model is employed for assessing measurement invariance or detecting DIF, then the restricted score specification appears to be most robust. Thus, the selection of the number of latent classes should only be based on this specification.

DeMars (2010) illustrate how significance tests based on the observed (raw) scores in reference and focal groups suffer from inflated type I error rates with an increased sample size if impact is present. This does not apply to the false alarm rate of Rasch mixture models because not a significance test but rather model selection via BIC is carried out. The rate of the BIC selecting the correct model increases with larger sample size if the true model is a Rasch mixture model. Since consistent estimates are employed, a larger sample size also speeds up convergence, which is particularly desirable for the saturated model if the number of latent classes and thus the number of parameters is high.

Given the correct number of classes, the different score models are all similarly suitable to detect the true classification if ability impact does not contain any additional information

about the DIF groups. However, if ability impact is highly correlated with DIF groups in the data and the ability groups thus coincide with the DIF groups, this information can be exploited by the unrestricted specifications while it distracts the restricted model.

Thus, while the selection of the number of latent classes should be based only on the restricted score specification, the unrestricted mean-variance and saturated specifications might still prove useful for estimating the Rasch mixture model (after $\hat{K}$ has been selected).

We therefore recommend a two-step approach for DIF detection via a Rasch mixture model. First, the number of latent classes is determined via the restricted score model. Second, if furthermore the estimation of the item difficulties is of interest, the full selection of score models can the be utilized. While the likelihood ratio test is not suitable to test for the number of latent classes, it can be used to establish the best fitting score model, given the number of latent classes. If this approach is applied to the full range of score models (saturated and mean-variance, both unrestricted and restricted), the nesting structure of the models needs to be kept in mind.

4. Empirical application: Verbal aggression

We use a dataset on verbal aggression (De Boeck and Wilson 2004) to illustrate this two-step approach of first assessing measurement invariance via a Rasch mixture model with a restricted score distribution and then employing all possible score models to find the best fitting estimation of the item difficulties.

Participants in this study are presented with one of two potentially frustrating situations (S1 and S2):

- S1: A bus fails to stop for me.
- S2: I miss a train because a clerk gave me faulty information.

and a verbally aggressive response (cursing, scolding, shouting). Combining each situation and response with either "I want to" or "I do" leads to the following 12 items:

S1WantCurse	S1DoCurse	S1WantScold	S1DoScold	S1WantShout	S1DoShout
S2WantCurse	S2DoCurse	S2WantScold	S2DoScold	S2WantShout	S2DoShout

First, we assess measurement invariance with regard to the whole instrument: we fit a Rasch mixture model with a restricted score distribution for $K = 1, 2, 3, 4$ and employ the BIC for model selection. Note that the restricted versions of the mean-variance and saturated model only differ in their log-likelihood by a constant factor and therefore lead to the same model selection. Results are presented in Table 2.

The BIC for a Rasch mixture model with more than one latent class is smaller than the BIC for a single Rasch model, thus indicating that measurement invariance is violated. The best fitting model has $\hat{K} = 3$ latent classes. Given this selection of K, we want to gain further insight in the data and thus want to establish the best fitting Rasch mixture model with $K = 3$ latent classes. Four different models are conceivable: either using a restricted or unrestricted score model, and either using a saturated or mean-variance specification. The results for all four options are presented in Table 3. Note that the models with restricted saturated score

Model	k	#Df	$\log L$	BIC
restricted (mean-variance)	1	13	−1900.9	3874.6
restricted (mean-variance)	2	25	−1853.8	3847.8
restricted (mean-variance)	**3**	**37**	**−1816.9**	**3841.4**
restricted (mean-variance)	4	49	−1793.0	3861

Table 2: DIF detection by selecting the number of latent classes $\hat{K}$ using the restricted Rasch mixture model.

Model	k	#Df	$\log L$	BIC
saturated	3	65	−1795.2	3955.1
restricted (saturated)	3	45	−1814.1	3880.6
mean-variance	3	41	−1812.2	3854.4
restricted (mean-variance)	**3**	**37**	**−1816.9**	**3841.4**

Table 3: Selection of the score distribution given the number of latent classes $\hat{K} = 3$.

distribution and restricted mean-variance score distribution lead to identical item parameter estimates. However, it is still of interest to fit them separately because each of the restricted specifications is nested within the corresponding unrestricted specification. Furthermore, the mean-variance distribution is nested within the saturated distribution.

As $K = 3$ is identical for all of these four models, standard likelihood ratio tests can be used for comparing all nested models with each other. Testing the most parsimonious score model, the restricted mean-variance model, against its unrestricted version and the restricted saturated model at a 5% level shows that a more flexible score model does not yield a significantly better fit. The p-value are 0.051 and 0.686, respectively. Hence, the restricted mean-variance distribution is adopted here which also has the lowest BIC.

To visualize how the three classes found in the data differ, the corresponding item profiles are shown in Figure 11.

- The latent class in the right panel (with 109 observations) shows a very regular zig-zag-pattern where for any type of verbally aggressive response actually "doing" the response is considered more extreme than just "wanting" to respond a certain way as represented by the higher item parameters for the second item, the "do-item", than the first item, the "want-item", of each pair. The three types of response (cursing, scolding, shouting) are considered increasingly aggressive, regardless of the situation (first six items vs. last six items).

- The latent class in the left panel (with 111 observations) distinguishes more strongly between the types of response. However, the relationship between wanting and doing is reversed for all responses except shouting. It is more difficult to agree to the item "I want to curse/scold" than to the corresponding item "I do curse/scold". This could be interpreted as generally more aggressive behavior where one is quick to react a certain way rather than just wanting to react that way. However, shouting is considered a very aggressive response, both in wanting and doing.

- The remaining latent class (with 53 observations considerably smaller), depicted in the

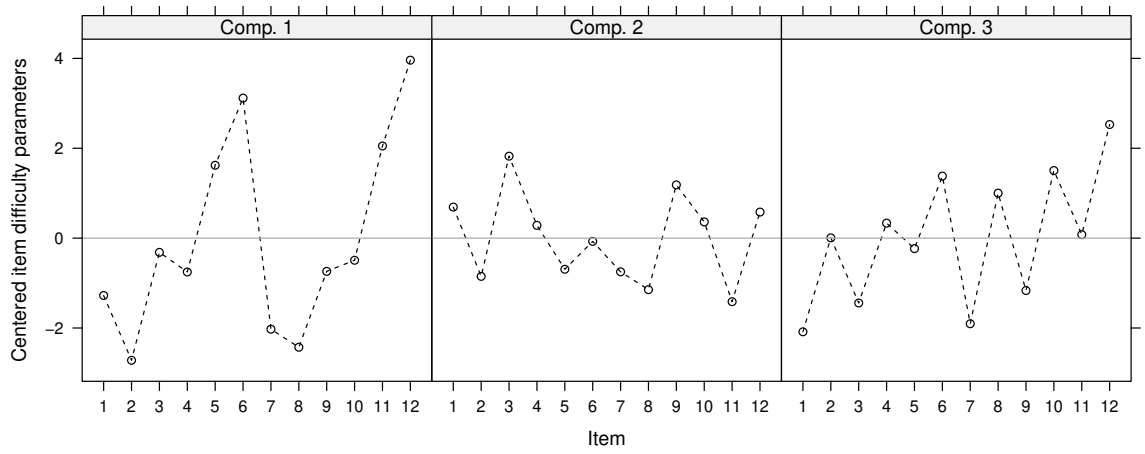

Figure 11: Item profiles for the Rasch mixture model with $\hat{K} = 3$ latent classes using a restricted mean-variance score distribution for the verbal aggression data.

middle panel, does not distinguish that clearly between response types, situations or wanting vs. doing.

Therefore, not just a single item or a small number of items have DIF but the underlying want/do relationship of the items is different across the three classes. This instrument thus works differently as a whole across classes.

In summary, the respondents in this study are not scalable to one single Rasch-scale but instead need several scales to represent them accurately. A Rasch mixture model with a restricted score distribution is used to estimate the number of latent classes. Given that number of classes, any type of score model is conceivable. Here, the various versions are all fairly similar and the restricted mean-variance specification is chosen based on likelihood ratio tests. Keep in mind that the resulting fits can be substantially different from each other as shown in the simulation study, in particular for the case of impact between DIF classes. The latent classes estimated here differ mainly in their perception of the type and the "want/do"-relationship of a verbally aggressive response.

5. Conclusion

Unlike in a single Rasch model, item parameter estimation is not independent of the score distribution in Rasch mixture models. The saturated and mean-variance specification of the score model are both well-established. A further option is the new restricted score specification introduced here. In the context of DIF detection, only the restricted score specification should be used as it prevents confounding effects of impact on DIF detection while exhibiting hit rates positively related to DIF effect size. Given the number of latent classes, it may be useful to fit the other score models as well, as they might improve estimation of group membership and therefore estimation of the item parameters. The best fitting model can be selected via the likelihood ratio test or an information criterion such as the BIC. This approach enhances the suitability of the Rasch mixture model as a tool for DIF detection as additional information

contained in the score distribution is only employed if it contributes to the estimation of latent classes based on measurement invariance.

Computational details

An implementation of all versions of the Rasch mixture model mentioned here is freely available under the General Public License in the R package **psychomix** from the Comprehensive R Archive Network. Accompanying the package at http://CRAN.R-project.org/package=psychomix is a dataset containing all the simulation results which were generated using R version 3.0.2, **psychomix** version 1.1-0, and **clv** version 0.3-2. This vignette was generated with R version 3.6.0, and **psychomix** version 1.1-7.

Acknowledgments

This work was supported by the Austrian Ministry of Science BMWF as part of the UniInfrastrukturprogramm of the Focal Point Scientific Computing at Universität Innsbruck.

References

Ackerman TA (1992). "A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective." *Journal of Educational Measurement*, **29**(1), 67–91.

Andersen EB (1972). "A Goodness of Fit Test for the Rasch Model." *Psychometrika*, **38**(1), 123–140.

Andrich D (1978). "A Rating Formulation for Ordered Response Categories." *Psychometrika*, **43**(4), 561–573.

Ankenmann RD, Witt EA, Dunbar SB (1999). "An Investigation of the Power of the Likelihood Ratio Goodness-of-Fit Statistic in Detecting Differential Item Functioning." *Journal of Educational Measurement*, **36**(4), 277–300.

Baghaei P, Carstensen CH (2013). "Fitting the Mixed Rasch Model to a Reading Comprehension Test: Identifying Reader Types." *Practical Assessment, Research & Evaluation*, **18**(5), 1–13.

Cohen AS, Bolt DM (2005). "A Mixture Model Analysis of Differential Item Functioning." *Journal of Educational Measurement*, **42**(2), 133–148.

De Boeck P, Wilson M (eds.) (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer-Verlag, New York.

DeMars CE (2010). "Type I Error Inflation for Detecting DIF in the Presence of Impact." *Educational and Psychological Measurement*, **70**(6), 961–972.

DeMars CE, Lau A (2011). "Differential Item Functioning Detection With Latent Classes: How Accurately Can We Detect Who Is Responding Differentially?" *Educational and Psychological Measurement*, **71**(4), 597–616.

Dempster A, Laird N, Rubin D (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B*, **39**(1), 1–38.

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

Frick H, Strobl C, Leisch F, Zeileis A (2012). "Flexible Rasch Mixture Models with Package **psychomix**." *Journal of Statistical Software*, **48**(7), 1–25. URL http://www.jstatsoft.org/v48/i07/.

Frick H, Strobl C, Zeileis A (2014). "Rasch Mixture Models for DIF Detection: A Comparison of Old and New Score Specifications." *Educational and Psychological Measurement*. doi:10.1177/0013164414536183. Forthcoming.

Gustafsson JE (1980). "Testing and Obtaining Fit of Data in the Rasch Model." *British Journal of Mathematical and Statistical Psychology*.

Holland PW, Thayer DT (1988). "Differential Item Performance and the Mantel-Haenszel procedure." In Wainer and Braun (1988), chapter 9, pp. 129–145.

Hong S, Min SY (2007). "Mixed Rasch Modeling of the Self-Rating Depression Scale: Incorporating Latent Class and Rasch Rating Scale Models." *Educational and Psychological Measurement*, **67**(2), 280–299.

Li F, Cohen AS, Kim SH, Cho SJ (2009). "Model Selection Methods for Mixture Dichotomous IRT Models." *Applied Psychological Measurement*, **33**(5), 353–373.

Li Y, Brooks GP, Johanson GA (2012). "Item Discrimination and Type I Error in the Detection of Differential Item Functioning." *Educational and Psychological Measurement*, **72**(5), 847–861.

Maij-de Meij AM, Kelderman H, van der Flier H (2010). "Improvement in Detection of Differential Item Functioning Using a Mixture Item Response Theory Model." *Multivariate Behavioral Research*, **45**(6), 975–999.

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

McLachlan G, Peel D (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

Molenaar IW (1995). "Estimation of Item Parameters." In Fischer and Molenaar (1995), chapter 3, pp. 39–51.

Nieweglowski L (2013). *clv: Cluster Validation Techniques*. R package version 0.3-2.1, URL http://CRAN.R-project.org/package=clv.

Preinerstorfer D, Formann AK (2011). "Parameter Recovery and Model Selection in Mixed Rasch Models." *British Journal of Mathematical and Statistical Psychology*, **65**(2), 251–262.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rand WM (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, **66**(336), 846–850.

Rasch G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* The University of Chicago Press.

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

Rost J (1991). "A Logistic Mixture Distribution Model for Polychotomous Item Responses." *British Journal of Mathematical and Statistical Psychology*, **44**(1), 75–92.

Rost J, von Davier M (1995). "Mixture Distribution Rasch Models." In Fischer and Molenaar (1995), chapter 14, pp. 257–268.

Schwarz G (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, **6**(2), 461–464.

Strobl C, Kopf J, Zeileis A (2014). "A New Method for Detecting Differential Item Functioning in the Rasch Model." *Psychometrika*. doi:10.1007/s11336-013-9388-3. Forthcoming.

von Davier M, Rost J (1995). "Polytomous Mixed Rasch Models." In Fischer and Molenaar (1995), chapter 20, pp. 371–379.

Wainer H, Braun HI (eds.) (1988). *Test Validity.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Zickar MJ, Gibby RE, Robie C (2004). "Uncovering Faking Samples in Applicant, Incumbent, and Experimental Data Sets: An Application of Mixed-Model Item Response Theory." *Organizational Research Methods*, **7**(2), 168–190.

Affiliation:

Hannah Frick, Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Hannah.Frick@uibk.ac.at, Achim.Zeileis@R-project.org
URL: http://eeecon.uibk.ac.at/~frick/, http://eeecon.uibk.ac.at/~zeileis/

Carolin Strobl
Department of Psychology
Universität Zürich
Binzmühlestr. 14
8050 Zürich, Switzerland
E-mail: Carolin.Strobl@psychologie.uzh.ch
URL: http://www.psychologie.uzh.ch/methoden.html