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

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 score distribution for the Rasch mixture model is introduced here which is restricted to be equal across latent classes. This causes 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 Rasch model and check for violations of measurement invariance such as differential item functioning (DIF). It 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 – here called Rasch mixture model to avoid confusion with mixed (effects) models and instead highlight its relation to mixture models – 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 manifest DIF tests, as 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. However, in a Rasch mixture model, the estimation of the item difficulties is not independent of this second aspect, 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, inflated type I error rates also occur in other DIF detection methods if impact is present, e.g., the Mantel-Haenszel and logistic regression precedures (Li, Brooks, and Johanson 2012; DeMars 2010).

Here, a simulations study is conducted to illustrate how Rasch mixture models react to impact, either alone or in combination with DIF. When using a Rasch mixture model for DIF detection, an influence of sole impact 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 Rasch mixture model specifically designed to detect DIF – regardless of whether or not impact is present and in which form.

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 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 $h(\cdot)$:

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

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

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

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:

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

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 $k = 1, \ldots, K$:

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

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

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

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. 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.

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

To obtain independence of the item parameter estimates from the specification of the score

distribution in the Rasch mixture model from Equation 4, we propose a novel specification of the score distribution in the Rasch mixture model. We suggest restricting the score parameters $\delta^{(k)}$ to be equal across the latent classes:

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

The independence of the item parameter estimates can then be seen easily from the definition of the posterior weights (Equation 5): $g(\cdot)$ can be moved out of the sum and then cancels out. 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. To complete the likelihood, a score distribution is simply fitted to the full-sample scores.

Consequently, using a restricted score specification, the likelihood of the Rasch mixture model (Equation 4) can be split into two factors: one depending on the general score parameters δ and the other one depending on latent class-specific prior probabilities $\pi^{(1)}, \ldots, \pi^{(K)}$ and item difficulties $\beta^{(1)}, \ldots, \beta^{(K)}$. The mixture is then only based on latent structure in the item difficulties, not on latent structure in both difficulties and scores. Also, the selection of the number of classes K is *not* affected by the specification of the score distribution $g(\cdot)$.

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.

Scenario	Latent class I		Latent class II	
	Abilities	Difficulties	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 ($\Delta > 0$), impact within	$\{-\Theta/2, +\Theta/2\}$	β^I	$\{-\Theta/2, +\Theta/2\}$	β^{II}
5 DIF ($\Delta > 0$), impact between	$\{-\Theta/2\}$	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.

To illustrate how suitable the different score specifications are to detect DIF (or lack thereof) a Monte Carlo study has been performed. The R system for statistical computing (R Core Team 2012) was used with the add-on packages **psychomix** (Frick *et al.* 2012) and **clv** (Nieweglowski. 2009).

3.1. Simulation design

The simulations included in Rost (1990) are the starting point for developing the design used here. Similar to the original simulation study, the item parameters represent a test 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 15 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. As in Rost (1990), the abilities are drawn from a discrete distribution. For simplicity, they are drawn with equal proportions from two values only,$-\Theta/2$ and $+\Theta/2$, thus creating a sample where half of the subjects have an ability of $-\Theta/2$ and the other half an abilty of $+\Theta/2$. Such a difference in ability is often called impact (Ackerman 1992).

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:

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

Drawing abilities from two normal distributions with means $-\Theta/2$ and $+\Theta/2$ leads to qualitatively similar results which are not reported here.

Impact and DIF, or lack thereof, can be combined in several ways. Table 1 provides an overview and Figures 1, 2, and 3 show illustrations:

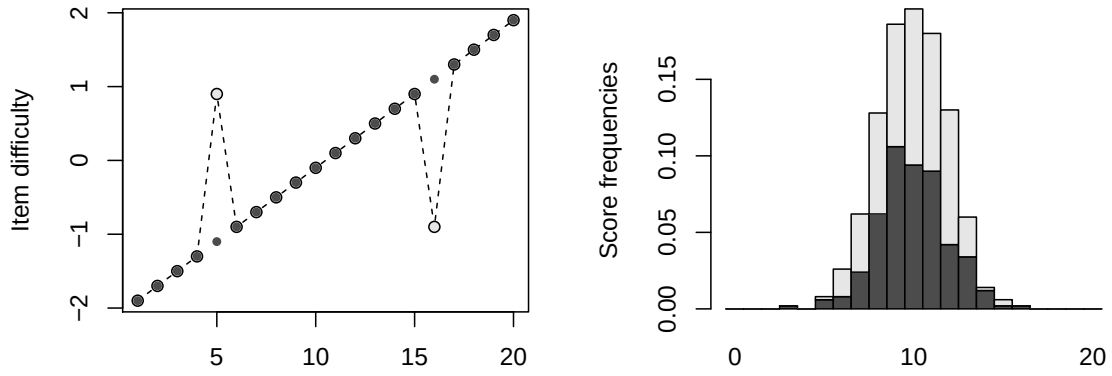

Figure 1: Scenario 2. Left: Item difficulties with DIF ($\Delta = 2$). Right: Stacked histogram of unimodal score distribution with homogeneous abilities ($\Theta = 0$).

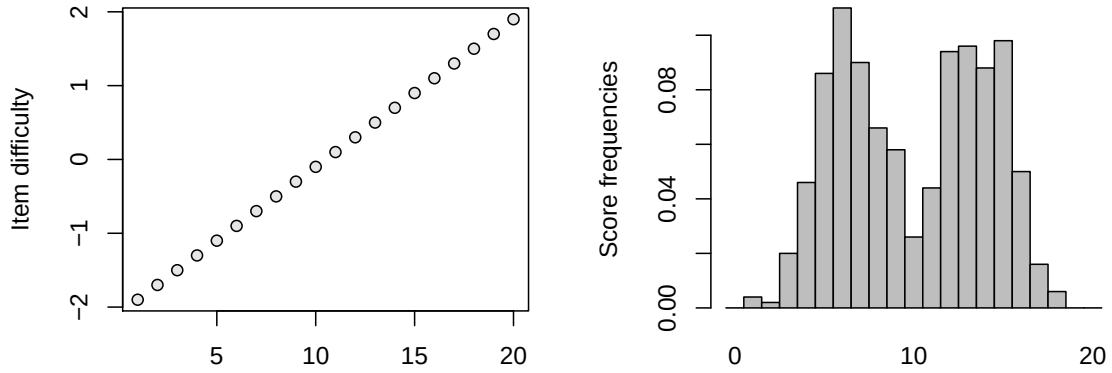

Figure 2: Scenario 3. Left: Item difficulties without DIF ($\Delta = 0$). Right: Histogram of bimodal score distribution with impact ($\Theta = 2$).

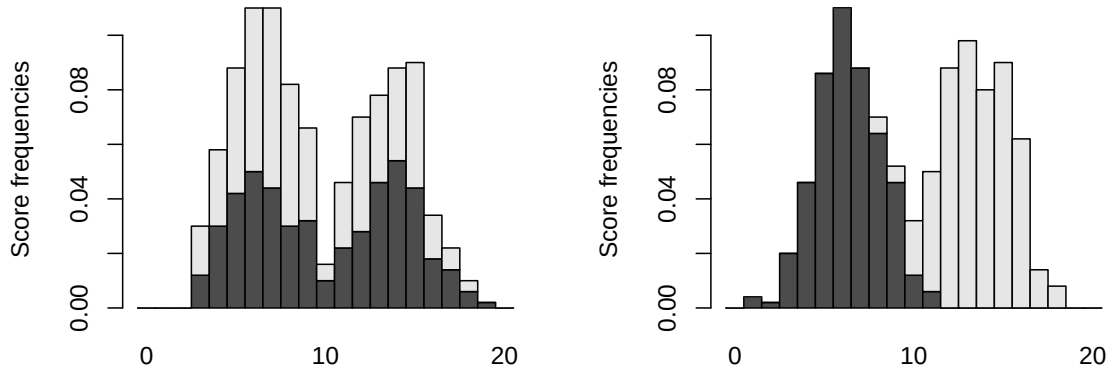

Figure 3: Stacked histograms of score distributions for Scenarios 4 (left) and 5 (right) with DIF ($\Delta = 2$). Left: impact within groups ($\Theta = 2$). Right: impact between groups ($\Theta = 2$). For item difficulties see Figure 1 (left).

- 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).
- 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 1.
- If the simulation parameter Θ for the impact is set to zero, abilities are homogeneous across all subjects (Scenarios 1 and 2) and the resulting score distribution is unimodal. For an illustration see the right panel of Figure 1. The histogram is shaded in light and dark gray for the two DIF groups present in this example from Scenario 2 and thus to be read like a "stacked histogram".
- If $\Theta > 0$, subject abilities are sampled from $\{-\Theta/2, +\Theta/2\}$ with equal weights, thus generating impact. When no DIF is included (Scenario 3), the resulting score distribution moves from being unimodal to being bimodal with increasing Θ. For the illustration in Figure 2 only a medium gray is used to shade the histogram as no DIF groups are present. However, if there is DIF (i.e., $\Delta > 0$), two combinations of DIF with impact are considered: Impact can occur within each DIF group (Scenario 4) or between DIF groups (Scenario 5). Illustrations of the resulting score distributions can be found in Figure 3.

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 within DIF groups* (Scenario 4) and the setting of *Impact between DIF groups* (Scenario 5) as a reference point.

Also note that the Scenarios 3 and 4 essentially correspond to the designs 1 and 2 of Rost (1990).

For each considered combination of Δ and Θ, 500 datasets of 500 observations each are generated. 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 $K = 1, 2, 3$.

3.2. Type I error and power in DIF detection

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 finding DIF in at least one item of the test. The empirical proportion among the 500 datasets with $\hat{K} > 1$ then essentially corresponds to the power of

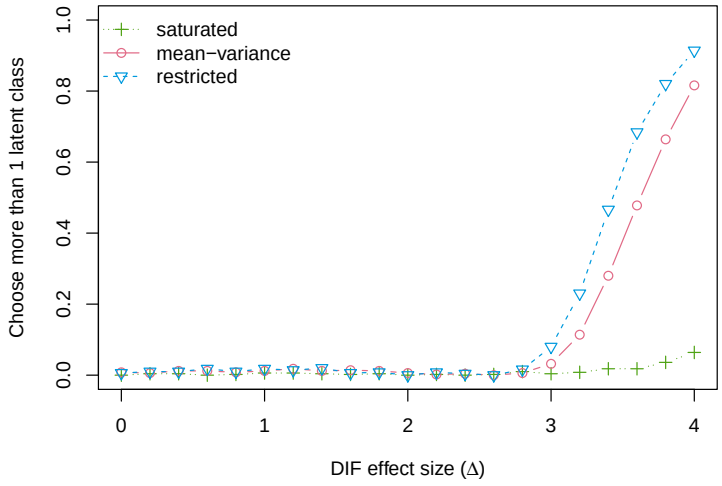

Figure 4: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 2 (DIF without impact, i.e., $\Theta = 0$).

DIF detection if $\Delta > 0$ (and thus two true latent classes) and to the associated type I error if $\Delta = 0$ (and thus only one true latent class).

Scenario 2: No impact with DIF

This scenario is investigated as it is 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 1) this corresponds to enlarging the spikes in the item profile.

Figure 4 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 corresponds to the type I error rate of DIF detection which is very close to zero for all three score distributions. With increasing $\Delta > 0$ the rate corresponds to the power of DIF detection and increases as well. For low values of Δ none of the three models is able to pick up the DIF but at around $\Delta = 3$ the two more parsimonious versions of the Rasch mixture model (with mean-variance and restricted score distribution) start to have increasing power which 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 saturated score model, however, has almost no power over the range of Δ considered. 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 almost all Rasch mixture models using a saturated score distribution.

Brief summary: The mean-variance and restricted model have a higher power than the saturated model in the absense of impact.

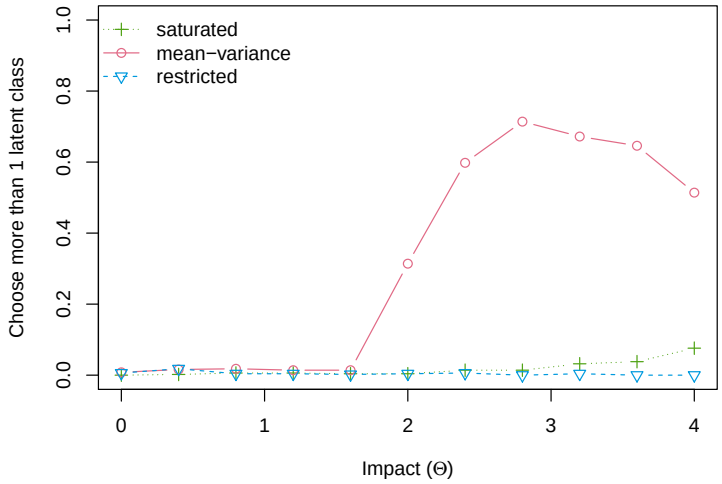

Figure 5: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 3 (impact without DIF, i.e., $\Delta = 0$).

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 2.

Here, the rate of choosing more than one latent class can be interpreted as a false discovery or false alarm rate (Figure 5). In the setting of a test this would correspond to the type I error of the test. The restricted score model is invariant against any latent structure in the score distribution and thus almost always suggests $\hat{K} = 1$ latent class based on the DIF-free item difficulties. 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 4 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 often suggests $\hat{K} > 1$ latent classes. 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.

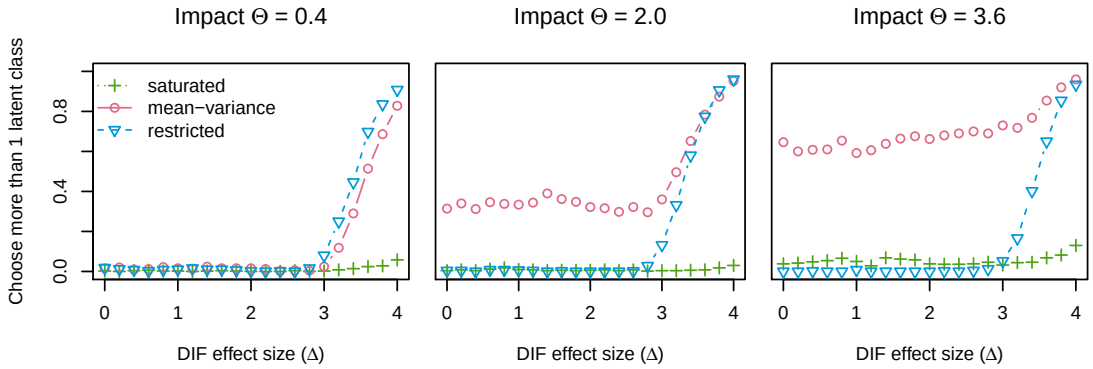

Figure 6: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 4 (impact within DIF groups).

Scenario 4: Impact within DIF groups

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 impact exists within each DIF group. Nonetheless, the score distribution is the same across both latent classes (illustrated in the left panel of Figure 3).

Figure 6 again shows the rate of choosing $\hat{K} > 1$ for increasing DIF effect size Δ for different levels of impact $\Theta = 0.4, 2.0, 3.6$. If impact is small (left panel with $\Theta = 0.4$), the rates are very similar to the case of completely homogeneous abilities without impact (Figure 4 with $\Theta = 0$). While the rates for the restricted and the saturated score model do not change substantially for an increased impact ($\Theta = 2.0$ and 3.6), the mean-variance model is influenced by this change in ability differences. While power is increased over the whole range of Δ, the type I error (or false alarm rate) at $\Delta = 0$ is increased to the same extent. Moreover, the detection rate only increases noticeably beyond the initial type I error rate at around $\Delta = 3$, i.e., the same DIF effect size at which the restricted and mean-variance specifications have power given homogeneous abilities without impact. Thus, given rather high impact ($\Theta = 3.6$) the power 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 5).

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 $K = 3$ classes. For the saturated and restricted specifications overselection is rare (occurring with rates of less than 5% or less than 1%, respectively). However, similar to Scenario 3 overselection is not rare for the mean-variance specification. Figure 7 depicts the rates of selecting $\hat{K} = 2$ and $\hat{K} = 3$ classes, respectively, for increasing Δ at $\Theta = 3.6$ (hollow symbols for Scenario 4). The rate for overselection ($\hat{K} = 3$) is already at around 30% for low values of Δ and even increases somewhat further starting from around $\Delta = 3$.

Brief summary: If impact is simulated within DIF groups, the mean-variance model has higher power than the saturated and resticted models. However, the latent classes estimated by the mean-variance model are mostly based on ability differences when the DIF effect size is

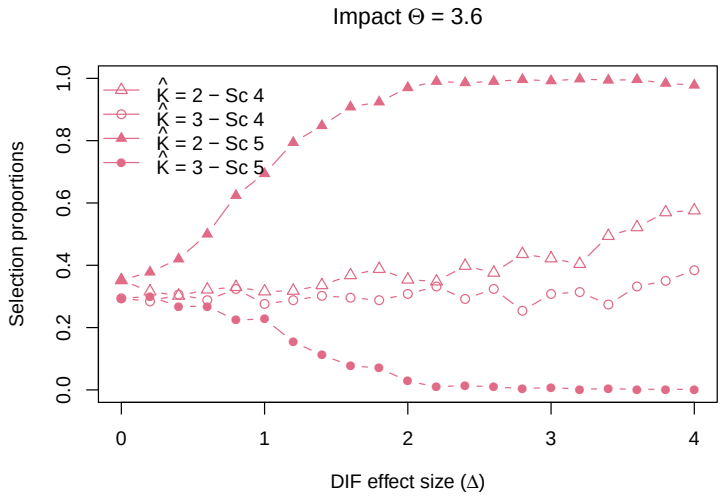

Figure 7: Rates of choosing the correct number of classes ($\hat{K} = 2$) or overselecting the number of classes ($\hat{K} = 3$) for the Rasch mixture model with mean-variance score specification in Scenarios 4 (hollow, impact within DIF groups) and 5 (solid, impact between DIF groups).

low. If the DIF effect size is high, the mean-variance model tends to overestimate the number of classes.

Scenario 5: Impact between DIF groups

In Scenario 5, there is also DIF (i.e., $\Delta > 0$) and impact. However, in contrast to Scenario 4 impact exists between the DIF groups but not within (see the right panel of Figure 3). 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 (compare left panel of Figure 8 and Figure 4). Both mean-variance and restricted specification have comparable power for DIF detection starting from around $\Delta = 3$ while the saturated specification has very lower power. As impact increases (middle and right panels of Figure 8), the power 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 this increased power is again coupled with a drastically increased false alarm rate at $\Delta = 0$. 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 in the case of impact within DIF groups (around $\Delta = 2$). The saturated model detects more than one latent class at a similarly low or lower rate than the other two models regardless of the level of impact.

Finally, the potential issue of overselection can be considered again. Figure 7 (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 25% of the datasets.

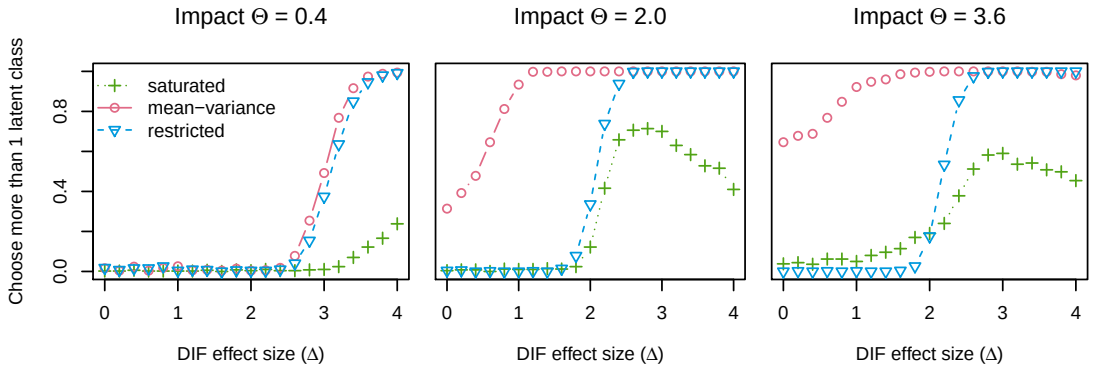

Figure 8: Rate of choosing a model with $\hat{K} > 1$ latent classes for data from Scenario 5 (impact between DIF groups).

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 power is 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.

3.3. Quality of estimation

Once the number of latent classes is established/estimated, it is of interest how well the estimated model fits the data. In the context of Rasch mixture models with different score distributions, the posterior probabilities $\hat{p}_{ik}$ (Equation 5) are crucial as the estimation of the item parameters depends on the score distribution only through these. Thus, if the $\hat{p}_{ik}$ were the same for all three score specifications, the estimated item difficulties were the same as well. Hence, it needs to be assessed 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.

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 and then 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). Figure 9 depicts the average Rand index for data from Scenario 4 (impact within DIF groups). 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

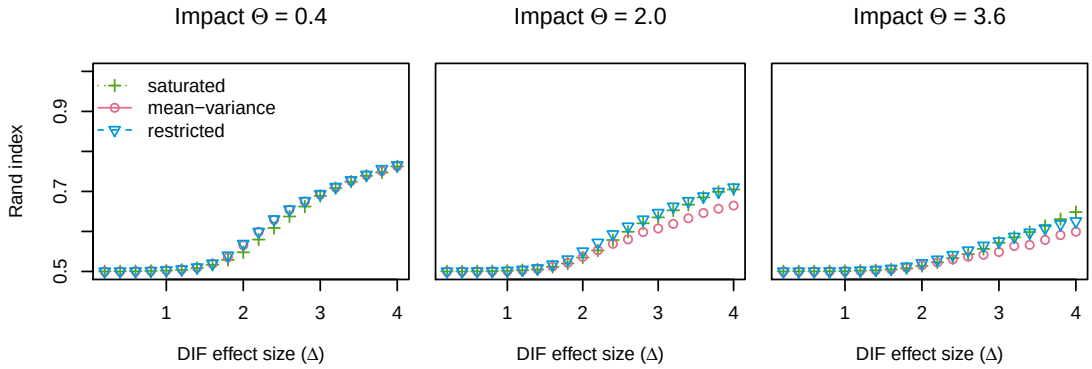

Figure 9: Average Rand index for models with $K = 2$ latent classes for data from Scenario 4 (impact within DIF groups).

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 – while it hardly ever chooses the correct number of classes in the first place.

However, in Scenario 5 where the score distribution contains information about the DIF groups, the three score specifications perform very differently as 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. However, the classifications obtained from the restricted score specification do not match the true groups well in this scenario. The reason is that the restricted score model is partially misspecified as the score distributions differ substantially across DIF groups.

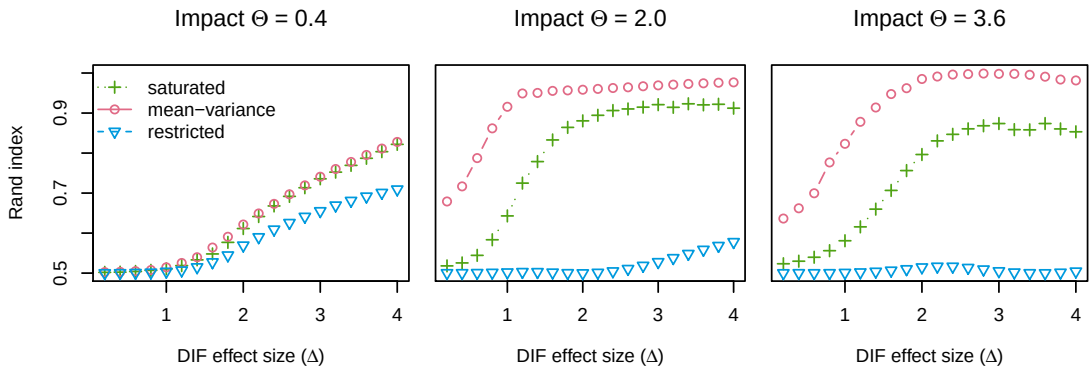

Figure 10: Average Rand index for models with $K = 2$ latent classes for data from Scenario 5 (impact between DIF groups).

3.4. 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 power than the other two specifications, i.e., violation of measurement invariance remains too often undetected.
- The mean-variance model has much better power. 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 power, comparable to the mean-variance model if abilities are rather homogeneous. But unlike the mean-variance specification, its type I error 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 be based on this specification.

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, 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 illustrate this approach on a dataset on verbal aggression (De Boeck and Wilson 2004). Participants 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 items:

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

First, we determine the number of latent classes K using the BIC for the Rasch mixture model with a restricted score specification:

Classes	1	2	3	4
BIC	3874.6	3847.8	3841.4	3865.5

Thus, $\hat{K} = 3$ latent classes are selected. Given this selection of K, four different models are conceivable: either using a restricted or unrestricted score model, and either using a saturated or mean-variance specification. Note that the models with restricted saturated score 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. The results for the verbal aggression data are shown in Figure 11. This shows that only the likelihood ratio test for restricted vs. unrestricted saturated specification is significant at 5% level while all other comparisons are (marginally) nonsignificant. Hence, the restricted mean-variance distribution is adopted here which also has the lowest BIC.

Figure 12 shows the corresponding item profiles.

- The latent class in the right panel (with 108 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 middle panel (with 112 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.

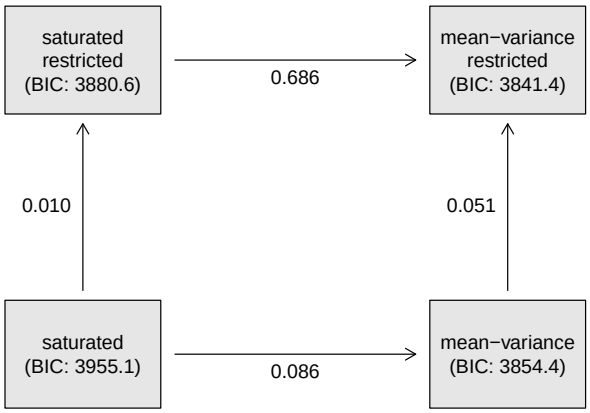

Figure 11: Likelihood ratio test p-values and BIC for Rasch mixture models with $K = 3$ latent classes and different score distribution specifications for the verbal aggression data. (Arrows denote the direction of nesting towards more restricted models.)

- The remaining latent class (with 53 observations considerably smaller), depicted in the left panel, does not distinguish that clearly between response types, situations or wanting vs. doing.

The respondents in this study are thus 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

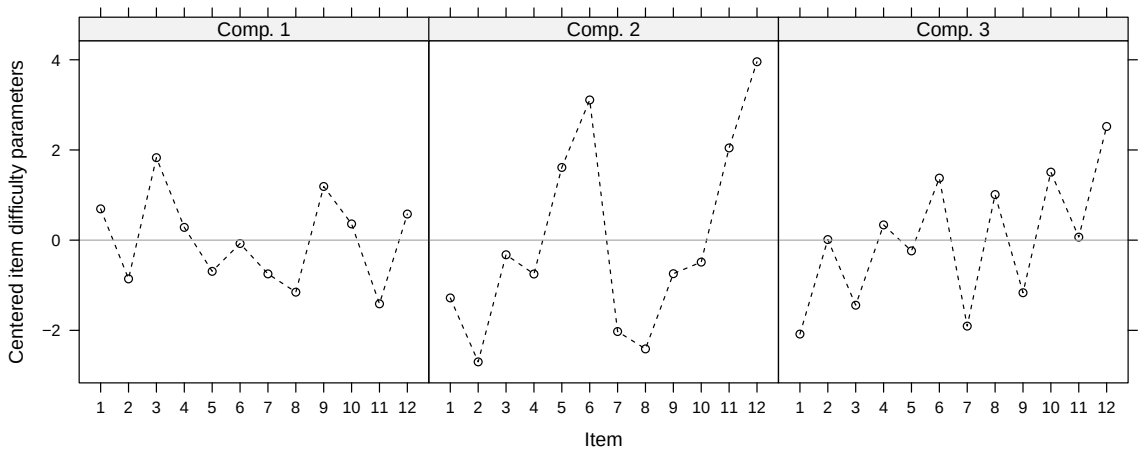

Figure 12: Item profiles for the Rasch mixture model with $\hat{K} = 3$ latent classes using a restricted mean-variance score distribution for the verbal aggression data.

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 preception 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 distribtution 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 testing for DIF, only the restricted score specification should be used as it prevents confounding effects of impact on DIF detection while exhibiting detection power 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 infomation 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 Pulic 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 vignette containing the simulation results and a replication of the verbal aggression example.

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.

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

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/.

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 (2009). ***clv**: Cluster Validation Techniques*. R package version 0.3-2, 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.

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.

R Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://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.

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.

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

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