Psychological Test and Assessment Modeling, Volume 60, 2018 (1), 101-139

Item response models for human ratings: Overview, estimation methods, and implementation in R

Alexander Robitzsch[1] *& Jan Steinfeld*[2]

Abstract

Item response theory (IRT) models for human ratings aim to represent item and rater characteristics by item and rater parameters. First, an overview of different IRT models (many-facet rater models, covariance structure models, and hierarchical rater models) is presented. Next, different estimation methods and their implementation in R software are discussed. Furthermore, suggestions on how to choose an appropriate rater model are made. Finally, the application of several rater models in R is illustrated by a sample dataset.

Keywords: multiple ratings, many-facet rater model, hierarchical rater model, R packages, parameter estimation, item response models

[1]Leibniz Institute for Science and Mathematics Education (IPN) at Kiel University, Kiel, Germany, and Centre for International Student Assessment, Germany. *Correspondence concerning this article should be addressed to:* Alexander Robitzsch, PhD, IPN, Olshausenstraße 62, D-24118 Kiel, Germany; email: robitzsch@ipn.uni-kiel.de

[2]Federal Ministry of Education, Science and Research, Austria

1 Introduction

Educational assessments often involve different approaches and procedures. Some abilities can be measured with closed answering formats such as multiple-choice questions, while other competencies, for example, expressive (productive) competencies, require constructed-response formats. One reason for why the latter are not so commonly used in large-scale assessments is that these kinds of tasks mostly require human judgment (rather than computer programs) to score answers or to assess their quality. Besides educational and language assessment, many other areas of testing require human judgment as well, such as the scoring of students within medical education programs (Tor & Steketee, 2011), the assessment of abilities using the approach of multiple mini-interviews (McLaughlin, Singer, & Cox, 2017), or large-scale placement tests (S. M. Wu & Tan, 2016). Therefore, possible rater effects must be taken into consideration.

Wind and Peterson (2018), who conducted a systematic review of the methods used in different application areas of rater studies, found that the research focus varies greatly. Some studies focus on the estimation of item difficulties, while others are more interested in the rating quality or the estimation of test-takers' ability. It is important to consider the main purpose of each study and to take into account the fact that the research focus may result in different study designs and that some estimation methods are superior to others. The research design and the estimation method chosen depend on the research question being investigated. Furthermore, the question of what kind of role the items and persons should have in the specific research should be considered.

First, items and persons could be seen as fixed, which means that each item and each person is associated with fixed model parameters, namely, an item difficulty and a person ability. As a consequence, the item responses X_{pi} of person p to item i are modeled as $P \times I$ independent random variables, given the fixed model parameters. Second, if persons and/or items are treated as random, this means that either persons and/or items are a random sample, it is necessary to make assumptions about the underlying distributions (see De Boeck, 2008). The consequences of the persons and/or items being treated as either fixed or random are that there is a change in the interpretation of the parameters and the resulting probabilities in the item response models (IRT models).

Performances are often graded by multiple raters in order to increase the reliability and objectivity of the ratings and to minimize rater errors (see Eckes, 2015 for a comprehensive overview). The expected degree of agreement (or nonagreement) depends on the attitudes and expectations of the raters, their knowledge, and the study design applied. For example, in studies in which raters grade performance more holistically (e.g., if there are no specific guidelines on how the raters should score the performance), a lower agreement is expected. When detailed scoring rules are applied, higher rater agreement can be expected. If detailed scoring rules are applied and broad training is provided for the raters, the ratings can be expected to be more homogeneous in terms of higher agreement between the performance scores. However, it could be expected that the application of detailed scoring rules yields ratings that are no longer locally independent, which is

a typical assumption made in IRT models (the residuals might correlate substantially; Wang, Su, & Qiu, 2014, see also Verhelst & Verstralen, 2001). The variable behavior of raters can be summarized under the label "rater effects". Depending on the knowledge of the raters, their attitudes, and their expectations of the performance, different raters may give different grades. Well-known rater effects include the effect of severity/leniency (Engelhard, 1992; Lunz, Wright, & Linacre, 1990), the halo effect (Bechger, Maris, & Hsiao, 2010; Myford & Wolfe, 2003), the central tendency, and the restriction of range of judgments (Engelhard, 1994; Saal, Downey, & Lahey, 1980).

Many different statistical approaches for analyzing multiple ratings are discussed in the literature (Eckes, 2017). To begin with, generalizability theory (G-theory; Brennan, 2001a) decomposes the total variance on a raw score metric (scores of raters on performance) into the additive variance components of the person, the items, and the raters. Both double and triple interactions (persons $\times$ items, persons $\times$ raters, and persons $\times$ items $\times$ raters) can be considered. G-theory treats items and persons as a sample of a theoretically infinite population of items and persons. The G-theory is useful regarding, for example, the formulation of rater effects, but it is also limited as the relationship of the components is treated as linear and additive in the raw score metric of items, which might not be appropriate.

In the context of the item response theory (IRT), several other methods have been proposed to model rater effects. These approaches are mostly based on the concept of virtual items, which are defined as the set of all combinations of original items and raters (see Rost & Langeheine, 1997). For example, in the case of two items and four raters, $2 \times 4 = 8$ virtual items can be created. A virtual item for a particular original item and a particular rater includes all ratings of the corresponding original item and rater, respectively. Based on virtual items, in the many-facet Rasch model (Linacre, 1989, 2017), the ratings of raters on all items and on all persons are decomposed into the additive effects of persons, items, and raters on the logit metric (more precisely, item $\times$ rater, or a matrix in which student essay $\times$ rater is shown). As illustrated in Figure 1a, each of the four raters rates two items. In total, there are two items and the responses to each of these two items are partitioned into four virtual items. The residuals among the virtual items are treated as being locally stochastically independent given a general person ability variable. A typical example of Figure 1a is the many-facet Rasch model, which results from the application of a restricted partial credit model to virtual items. Systematic differences in rater behavior are modeled by allowing item difficulties to differ between raters. However, the ratings that correspond to generalized items are assumed to be locally stochastically independent. This assumption is typically violated in many applications because the ratings of one single item by two raters will appear to be more similar than the ratings of two different items by two raters. Therefore, additional person-item interaction effects have to be considered.

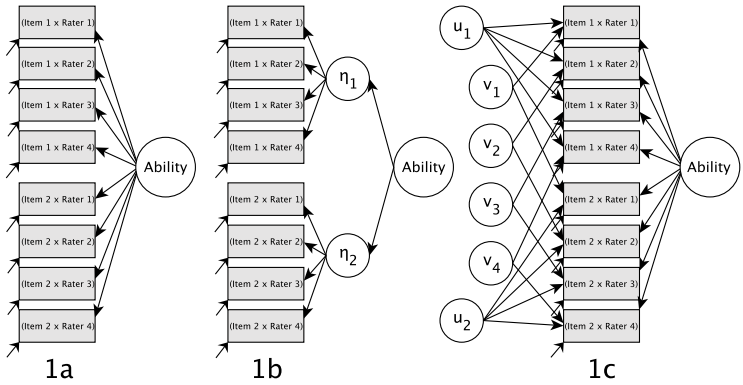

Figure 1:

All models 1a, 1b, and 1c represent eight virtual items, where each rater rated two of the virtual items. In both models 1a and 1b, the ratings were locally independent, whereas in model 1c, the additional parameters u and v were introduced to account for the interaction between persons and items as well as between persons and raters. Model 1a depicts the many-facet rater model, model 1b the hierarchical rater model, and model 1c the generalized many-facet rater model.

In Figure 1, the additional dependence caused by rating the same item is taken into account by a hierarchical rater model (Patz, Junker, Johnson, & Mariano, 2002; DeCarlo, 2005; DeCarlo, Kim, & Johnson, 2011). Person ability causes true ratings η of the two items, which are themselves measured by 2×4 observed ratings (i.e., the virtual items). Moreover, it is possible that the rating of a particular rater on the first item influences the rating on the second item (halo effect). In this case, additional dependence is introduced and the local independence assumption in Figure 1b is violated. In the generalized many-facet rater model depicted in Figure 1c, person-item and person-rater interactions are modeled by additional random effects (latent variables; Wang et al., 2014) that capture the violation of local independence in Figures 1a und 1b. It should be noted that local dependence can be alternatively represented as correlated residuals in Figure 1c. In the next section, these three different modeling approaches are formally described and are introduced as special cases of IRT models applied to the polytomous item responses of virtual items.

2 Item Response Models for human raters

In the following section, different item response models for human ratings are introduced. First, an overview of IRT models is presented. Then, these IRT models are extended to include rater effects for modeling rating data for human raters. In particular, we distinguish between the approaches of many-facet rater models, covariance structure models, and hierarchical rater models.

2.1 Item response models for polytomous data

Here, we provide a short review of the most frequently used IRT models for polytomous data. With X_{pi} we denote the polytomous item response of person p to item i. While the items are often treated as fixed, person parameters are often assumed to be random (see Holland, 1990) and are modeled by a distribution (e.g., a normal distribution or located latent classes). In the following description, we will mostly choose a unidimensional distribution of the ability (latent trait) θ_p, although the extension to multidimensional traits does not substantially change the interpretation of the models.

Partial credit model

The partial credit model (PCM; Masters, 1982) is an item response model for two or more ordered categories. The item response probability for responding to category $k = 0, \ldots, K_i$ is given as

$$P(X_{pi} = k|\theta_p) \propto \exp\{k\theta_p - b_{ik}\} \tag{1}$$

The symbol $\propto$ means that the right-hand side of Equation (1) sums to one across all categories k. The model has the property that persons with high abilities θ_p tend to respond in high categories k. The parameter b_{ik} indicates an item-category-specific intercept. This parameter is also often reparameterized in the form $b_{ik} = k\beta_i - \sum_{h=0}^{k} \tau_{ih}$ with a general item difficulty β_i and item thresholds τ_{ih}. The PCM belongs to the family of Rasch models and shares the important properties of the Rasch model that the sum score $S_p = \sum_i X_{pi}$ is a sufficient statistic for the person parameter θ_p and the person and item parameters are separable (Andersen, 1980). Therefore, conditional maximum likelihood estimation can be used as an estimation approach that provides item parameter estimates without the need to specify the ability distribution (see Section 3). A restricted form of the PCM is the linear logistic test model (LLTM; Fischer, 1973), which models the item-specific intercepts as a linear function of basis parameters and is given as

$$b_{ik} = \sum\nolimits_{m=1}^{M} q_{ikm}\gamma_m \tag{2}$$

where γ_m are basis item parameters and q_{ikm} are known prespecified values. Specific hypotheses can be tested by imposing restrictions on the PCM in Equation (1). For example, a rating scale model (Andrich, 1978) can be formulated as a particular LLTM, in which the model has item difficulty parameters and item thresholds that are assumed to be invariant across items.

Generalized partial credit model

The generalized partial credit model (GPCM) is a generalization of the PCM and was introduced by Muraki (1992). This model includes an additional item-specific discrimination parameter a_i and allows the items to have different reliabilities. It is formulated as

$$P(X_{pi} = k|\theta_p) \propto \exp\{ka_i\theta_p - b_{ik}\} \tag{3}$$

In most applications, the GPCM provides a better model fit than the PCM. Items with larger item discriminations are preferred because they are more informative in discriminating between persons with lower and higher ability values. As in the PCM, item discriminations a_i as well as item-category intercepts b_{ik} can be modeled as linear functions of the basis item parameters (Embretson, 1999); this makes the estimation of more parsimonious models possible. It should be emphasized that the weighted sum score $S_p = \sum_i a_i X_{pi}$ is a sufficient statistic for the person parameter θ_p.

Graded response model

The graded response model (GRM) proposed by Samejima (1969) belongs to the class of so-called cumulative IRT models. The item response probabilities are given as

$$P(X_{pi} = k|\theta_p) = G(a_i\theta_p - b_{i,k+1}) - G(a_i\theta_p - b_{ik}) \qquad (b_{i0} = 0, b_{i,K_i+1} = \infty) \quad (4)$$

where G is a link function that is typically the logistic link function or the probit link function. The model includes item discriminations a_i and ordered item intercepts b_{ik}. It is often found that the GRM and the GPCM provide similar fit to empirical datasets (Forero & Maydeu-Olivares, 2009) and, hence, there are no crucial consequences of choosing one of the two models. Again, the item parameters can be formulated as linear functions to estimate restricted versions of the GRM. For the probit link function, Equation (4) can be rewritten as $X^*_{pi} = a_i\theta_p + e_{pi}$ where X^*_{pi} is an underlying continuous variable for the ordinal item X_{pi} and e_{pi} is a standard, normally distributed residual. The ordinal item X_{pi} is obtained by discretizing the continuous variable X^*_{pi} with respect to thresholds b_{ik}. Using the variable X^*_{pi} has the advantage that correlated residuals can be specified in the GRM, which can model violations of local independence. However, in this situation, marginal maximum likelihood estimation is no longer computationally feasible and limited information estimation procedures have to be applied (see Section 3).

Covariance structure model

The normal distribution is probably the most frequently applied distribution. Sometimes the question arises whether the normal distribution can also be applied to ordinal items. However, the probability density of the normal distribution is defined on the real line and not on discrete values. Therefore, a misspecified model results if the normal distribution is applied to ordinal items. The assumed normal density is given as

$$f(X_{pi} = k|\theta_p) = (2\pi\sigma_i)^{-\frac{1}{2}} \exp\Big\{ - \frac{(k - a_i\theta_p - \mu_i)^2}{2\sigma_i^2} \Big\}, \text{ i.e. } X_{pi} = \mu_i + a_i\theta_p + e_{pi} \quad (5)$$

An item i is parameterized with an item mean μ_i, an item discrimination a_i, and a residual variance $\sigma_i^2 = Var(e_{pi})$. Unfortunately, the item parameters of the GPCM or the GRM cannot be simply converted into the parameters of the normal distribution in Equation (5). However, in some applications, the item and distribution parameters

from the covariance structure model (CSM; often referred to as confirmatory factor analysis) shown in Equation (5) can be more easily interpreted than the parameters of the GPCM or GRM. More formally, in a CSM, the mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\gamma})$ of the I items $X_{p1}, \ldots, X_{pI}$ and the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\gamma})$ are modeled as functions of an unknown parameter vector $\boldsymbol{\gamma}$ (Bollen, 1989). In a CSM, the covariance matrix is represented as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$, where $\boldsymbol{\Lambda}$ is the loading matrix, $\boldsymbol{\Phi}$ is the factor covariance matrix and $\boldsymbol{\Psi}$ is the residual covariance matrix. Then, the vector $\boldsymbol{\gamma}$ contains elements of the mean vector, loadings and elements of the factor covariance, and residual covariance matrices. When applied to ordinal data, the CSM is a so-called pseudo-likelihood estimation approach as the assumed likelihood function is misspecified (Yuan & Schuster, 2013). Interestingly, Arminger and Schoenberg (1989) showed that the mean structure and the covariance structure in Equation (5) can be consistently estimated in a confirmatory factor analysis based on a misspecified normal distribution for ordinal items. However, so-called maximum likelihood robust standard errors should be used, in order to ensure that valid statistical inferences can be made in the case of a misspecified likelihood (White, 1982). Alternatively, the bootstrap resampling method of persons can be used to obtain valid standard errors (Berk et al., 2014).

2.2 Many-facet rater model

The IRT models presented in the following paragraphs are based on virtual items of every combination of an item and a rater (see Figure 1). We denote the corresponding item responses as X_{pir} for person p to item i rated by rater r. Unidimensional many-facet rater models (MFRM) are obtained by applying the PCM, the GPCM, or the GRM to these virtual items. The item response probability in the extension of the GPCM is given as

$$P(X_{pir} = k|\theta_p) \propto \exp\{ka_{ir}\theta_p - b_{ikr}\} \tag{6}$$

and, for the GRM, it is written as

$$P(X_{pir} = k|\theta_p) = G(a_{ir}\theta_p - b_{ir,k+1}) - G(a_{ir}\theta_p - b_{irk}) \tag{7}$$

Typically, constrained versions of these models are applied to rating data. In the family of Rasch models, the item discriminations a_ir in the GPCM (Equation 6) are all set to one (also labeled as Rasch-MFRM in the following). A Rasch-MFRM (Linacre, 1989) imposes additional restrictions on item parameters such that

$$P(X_{pir} = k|\theta_p) \propto \exp\{k\theta_p - k\beta_i - k\alpha_r - \sum\nolimits_{h=0}^{k} \tau_{ih}\} \tag{8}$$

In this specification, the parameter β_i refers to the general item difficulty, α_r is the rater severity parameter and τ_{ih} are item-step parameters. The model specified in Equation (8) is a particular LLTM of the PCM applied to virtual items with linear constraints on item-category intercepts b_{irk}. It should be emphasized that rater effects are assumed to be homogeneous across all items in Equation (8). An important extension to Equation

(8) is the introduction of further interaction effects between items and raters, which allows for systematic item-specific rating behavior. The more restrictive model with homogeneous rater effects can be tested against the more complex model that allows for item-rater interactions. Rater centrality/extremity (see Wolfe, 2014) can be modeled by including rater-step parameters α_{rh} in Equation (8). We note that an identification condition has to be assumed in order to estimate (8) (e.g., $\Sigma_r \alpha_r = 0$).

The Rasch-MFRM has the advantage that the unweighted sum score $S_p = \Sigma_{ir} X_{pir}$ is a sufficient statistic for the person parameter θ_p. By using the many-facet Rasch model as a scaling model for obtaining person parameter estimates, an implicit decision about an equal weighting of items is made. From the perspective of item fit in real data applications, items as well as raters will typically assess the person ability with different precision. Therefore, an item response model that includes discrimination parameters will almost always result in better model fit. Besides severity-leniency effects or scale-usage effects of raters, raters can also differ in the reliability of the ratings they provide. The item-rater discrimination parameter $a_i r$ is a measure of the reliability of the ratings of item i and rater r (M. Wu, 2017). A more parsimonious model, which can also often be useful, linearly decomposes the item-rater discrimination, such that $a_{ir} = a_i + a_r$. Submodels that include only item discriminations ($a_{ir} = a_i$) or only rater discriminations ($a_{ir} = a_r$) provide further interesting diagnostic tools for studying the behavior of items and raters.

We emphasize that the GPCM (6) and the GRM (7) are often specified in a restricted form in which item-rater parameters follow a linear function, such as in the LLTM. These models are implemented in the R packages discussed in Section 3 of this paper.

2.3 Generalized many-facet rater model

As argued in the introduction, ratings are not locally independent across items and raters. First, different raters evaluate the performance of a student on an item, which typically introduces some dependency because additional item-specific factors besides general ability are at play. Hence, an additional student-item interaction effect has to be modeled. Second, the rating of one item by a rater can also influence the rating of another item by the same rater (the halo effect). Therefore, an additional student-rater interaction needs to be modeled. The MFRM can be extended to include these two additional random effect parameters u_{pi} and v_{pr} to model local dependence. The resulting generalized many-facet rater model (GMFRM; Wang et al., 2014; Verhelst & Verstralen, 2001, for a version for dichotomous ratings) can be written as

$$P(X_{pir} = k|\theta_p, u_{pi}, v_{pr}) \propto \exp\{k\alpha_i\theta_p + ku_{pi} + kv_{pr} - k\beta_i - k\alpha_r - \sum\nolimits_{h=0}^{k} \tau_{ih}\} \quad (9)$$

Several submodels of (9) can be estimated. A version of (9) that sets all item discriminations a_i to one is a multidimensional Rasch model (Wang et al., 2014) with random person-item and person-rater effects. The size of the variance components $\sigma_i^2 = Var(u_{pi})$ and $\sigma_r^2 = Var(v_{pr})$ quantifies the degree of the dependency of the

ratings. In some applications, it seems useful to include only the item or rater random effect for local dependence. The GMFRM models the additional dependence caused by ratings of the same items and by the same raters as additional random effects that prevent the assumption of local independence. Alternatively, the GPCM (9) can be substituted by a GRM using a latent variable representation. Using this approach, the random effects can be integrated out so that only person ability appears as a person variable in the model (Tuerlinckx & De Boeck, 2004; see also Ip, 2010). However, this equivalent model introduces correlated residuals, as rating variables X_{pir} for the same item i and for the same rater r are typically positively correlated. It must be emphasized that moving from the model with random effects to the equivalent model with correlated residuals implies a change in the metric of item parameters because the ability metric has changed. More formally, integrating out the random effects u_{pi} and v_{pr} from (9) results in the conditional response probability

$$P(X_{pir}=k|\theta_p) \propto \exp\{k\lambda_{ir}\alpha_i\theta_p - k\lambda_{ir}\beta_i - k\lambda_{ir}\alpha_r - \sum\nolimits_{h=0}^{k}\lambda_{ir}\tau_{ih} \qquad (10)$$
$$\text{with} \quad \lambda_{ir} = (\delta^2\sigma_i^2 + \delta^2\alpha_r^2 + 1)^{\frac{1}{2}}\}$$

where $\delta = 0.583$ is a positive constant (see Ip, 2010). As the multiplication factor λ_{ir} is always smaller than one, all item parameters are shrunken to the extent of local dependence caused by person-item and person-rater interactions. Hence, comparisons of the item parameters of the GMFRM and the MFRM should consider the transformation formula in Equation (10) for item parameters. The size of the residual correlations in (10) can also be computed based on the variance components of the random effects in model (9).

2.4 Covariance structure model and generalizability theory

Instead of modeling the ordinal virtual items of the rating data with an item response model for polytomous item responses, a CSM can alternatively be applied using normal distributions for modeling the virtual items X_{pir}. The mean structure can be represented by general item effects and general rater effects for modeling severity. The covariance structure can be modeled as a confirmatory factor model $\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}^T + \mathbf{\Psi}$ in which the distribution parameters of person ability are represented in the covariance matrix $\mathbf{\Phi}$ of latent factors. Violations of local independence caused by ratings of the same items and the same raters can be specified either as additional factors appearing in the covariance matrix $\mathbf{\Phi}$ or as a patterned residual covariance matrix $\mathbf{\Psi}$. As argued above, the CSM provides consistent estimates of the mean and covariance structure for ordinal items with misspecified normal distribution likelihood (Arminger & Schoenberg, 1989). This also holds true if the statistical models of G-theory (Brennan, 2001a) are applied to ordinal items because these models are particular cases of CSMs.

2.5 Hierarchical rater model

The GFRM and the CSM model the dependency caused by rating the same items by including an additional random effect or correlated residuals. Hierarchical rater models (HRM; Patz et al., 2002; DeCarlo et al., 2011) assume the existence of a discrete true rating η_{pi} of a person p on an item i. However, the true rating is not observed; rather, it is only indirectly measured by the ratings of several raters. The true rating categories of all items serve as indicators of the person ability θ_p. As a consequence, the item response ratings X_{pir} are hierarchically modeled, given true items η_{pi}, which are also hierarchically modeled, given the person ability θ_p. At the first level, a probability distribution $P(X_{pir} = k|\eta_{pi})$ specifies a rater model, while at the second level, the distribution $P(\eta_{pi} = \eta|\theta_p)$ is specified. At the second level, the GPCM can be chosen for modeling true ratings and can be written as

$$P(\eta_{pi} = \eta|\theta_p) \propto \exp\{\eta\alpha_i\theta_p - b_{ik}\} \tag{11}$$

For the rater model at the first level, two different model specifications have been proposed in the literature. Patz et al. (2002) used a discretized normal distribution as the rater model in the originally proposed hierarchical rater model (HRM; see also Casabianca & Wolfe, 2017):

$$P(X_{pir} = k|\eta_{pi}) \propto \exp\Big(-\frac{1}{2\psi_{ir}^2}[k - (\eta_{pi} + \phi_{ir})]^2\Big) \tag{12}$$

The parameter ϕ_{ir} represents a rater severity parameter that models the systematic displacement of the ratings of rater r from the true rating η_{pi}. The variance parameter ψ_{ir} is a measure of the reliability of the rater. Large values for the variance represent a high precision of the rater. The parameters ϕ_{ir} and ψ_{ir} can also be assumed to be invariant across items if a more parsimonious model should be estimated. We want to emphasize that (11) only parameterizes rater severity and rater imprecision. As noted by Patz et al. (2002), the estimation of severities ϕ_{ir} poses computational challenges for small rater-variances ψ_{ir}.

DeCarlo et al. (2011) proposed a hierarchical rater model based on a latent class signal detection model (HRM-SDT) in which the different scale usage of the raters can also be modeled. The item response probabilities in the rater model are specified as a GRM:

$$P(X_{pir} = k|\eta_{pi}) = G(d_{ir}\eta_{pi} - c_{ir,k+1}) - G(d_{ir}\eta_{pi} - c_{irk}) \tag{13}$$

where d_{ir} are item-rater discriminations and c_{irk} are item-rater-category thresholds. Large values for d_{ir} represent highly discriminating raters. Rater severity/leniency or rater centrality/extremity is represented by different values of the thresholds c_{irk}. Ideal raters, who always agree with the true rating category η, have very large discriminations d_{ir} (e.g., larger than 100) and item thresholds are given as $c_{irk} = d_{ir} \times (k - 0.5)$. It is evident that both hierarchical rater models take the dependence caused by rating the same items into account. The HRM-SDT of DeCarlo et al. (2011) appears to be more

flexible in modeling different rater behavior than the HRM of Patz et al. (2002), although it is possibly more difficult to estimate when only a small amount of data is available. However, neither model takes into account the additional dependence structure that occurs when multiple items are rated by one rater. If halo effects exist in applications, the GMFRM or a model with correlated residuals could be used. Alternatively, the hierarchical rater model can be extended to include an additional dependence structure (see also Wang et al., 2014) or random person-rater effects.

3 Estimation methods and their implementation in R packages

In this section, we present a brief overview of estimation methods that can be used for the rater models introduced in Section 2. We focus on the implementation of these methods in a number of recently released R packages (R Core Team, 2018) written by the authors (**immer**, Robitzsch & Steinfeld, 2018; **TAM**, Robitzsch, Kiefer, & Wu, 2018; **sirt**, Robitzsch, 2018b; **LAM**, Robitzsch, 2018a). This focus is intended to provide a basis for the illustrative examples discussed later; it does not imply general recommendations for real data analyses (see Rusch, Mair, & Hatzinger, 2013, for a more comprehensive overview of R packages for IRT). In general, two broad classes can be distinguished: maximum likelihood (ML) and Bayesian estimation. Several variants of ML estimation are discussed (see also Holland, 1990).

Marginal maximum likelihood (MML) estimation (also labeled as full information maximum likelihood estimation, FIML) estimates model parameters under a distributional assumption about person ability (and further random effects). In most cases, the normal distribution is chosen for person ability. As person ability is a latent variable, it is integrated out in the likelihood that the estimation problem can essentially be reduced to estimating item parameters (and rater parameters) and person distribution parameters (means, variances, and covariances). Essentially, MML operates under the assumption of random persons. Therefore, a person distribution is described by a statistical model and each person is not treated as a fixed entity for which the item response model holds. The expectation maximization (EM) algorithm is often employed for MML estimation (Aitkin, 2016). MML estimation for the Rasch-MFRM is available in the function `TAM::tam.mml.mfr()` of the **TAM** package. Several submodels of the MFRM that allow for different item discriminations can be estimated with `TAM::tam.mml.2pl()` or `sirt::rm.facets()`. An MML implementation of the HRM-SDT model of DeCarlo et al. (2011) can be found in `sirt::rm.sdt()`. In principle, the HRM of Patz et al. (2002) can also be estimated with the MML method, although an implementation is available in any of the R packages discussed in this section. MML estimation for CSMs based on a multivariate normal distribution can be found in the **lavaan** package (Rosseel, 2012) or in the `LAM::mlnormal()` function. G-theory models have equal linear discrimination parameters and fall into the class of linear mixed effects models that can be estimated with the **lme4** package (Bates, Mächler, Bolker, & Walker, 2015).

In joint maximum likelihood (JML) estimation (Lord, 1980; also labeled as fixed effects estimation), person parameters and item parameters are estimated simultaneously. Essentially, persons are treated as fixed and a single parameter is estimated for each person. Hence, no distributional assumption of person ability is needed. The JML estimation is only computationally stable for Rasch-MFRMs and is implemented in the Facets software (Linacre, 1989, 2017). JML has the disadvantage that the number of estimated parameters increases with the number of persons in the sample, which induces the well-known bias in JML estimation (Andersen, 1980). For the PCM, a simple bias-correction formula has been proposed (Andersen, 1980). However, this formula cannot be easily generalized to rating data with complex rating designs in which the number of ratings per person and per item differs. Considering the critique of JML in most of the psychometric literature, resampling methods and analytical methods (Hahn & Newey, 2004) have been proposed, which practically remove the bias caused by JML estimation. Bertoli-Barsotti, Lando, and Punzo (2014) proposed a modification to the likelihood function of the Rasch model for JML estimation that removes most of the bias in item parameters. The reason for the JML bias is that there is no simple way to handle persons with extreme scores (persons score in the lowest category or in the largest category for all items). The so-called ϵ-adjustment method of Bertoli-Barsotti et al. (2014) essentially applies a linear function to the sum score $S_p = \Sigma_i X_{pi}$ in order to map the interval $[0, M_p]$ (M_p is the maximum score for person p) onto $[\epsilon, M_p - \epsilon]$. It should be emphasized that all scores are linearly transformed. The ϵ-adjustment approach is implemented in the `immer::immer_jml()` function of the **immer** package (Robitzsch & Steinfeld, 2018) and extends the method of Bertoli-Barsotti et al. (2014) to polytomous item responses and multiple-matrix designs with arbitrary missing patterns. Therefore, this JML estimation method with bias-correction enables the estimation of the Rasch-MFRM. The statistical properties of the parameter estimates can be seen as being superior to alternative JML implementations of the Rasch-MFRM (for example, in the Facets software; Linacre, 2017). Depending on the application, JML can be substantially faster than MML estimation and, hence, JML could be seen as a viable estimation alternative even if persons are treated as random.

Conditional maximum likelihood (CML; Andersen, 1980) estimation also avoids a specification of the distribution of person ability as person parameters are completely removed in the estimation approach. Hence, CML can be used under the perspective of random persons as well as fixed persons. CML can only be applied for Rasch-MFRMs. The basic idea of CML is that the likelihood of a particular item response pattern with sum score v is conditioned on the sum of the likelihoods of all response patterns with sum score v. It can be shown that the corresponding ratio is independent of person ability and that CML provides consistent item parameter estimates (like MML estimates; van der Linden, 1994). It should be emphasized that CML becomes cumbersome in rating designs in which not all persons are rated by the same items and the same raters because the CML computations must be separately evaluated for every missing data pattern. CML for Rasch-MFRMs is available in the **eRm** package (Mair & Hatzinger,

2007) and the `immer::immer_cml()` function.

MML and CML estimation can be computationally demanding with complex rating data designs because there can be a large number of virtual items with many missing values. To reduce the computational burden, so-called limited information estimation approaches have been proposed, which do not rely on modeling the full item response patterns but, rather, operate on the aggregated information of the data.

The diagonally weighted least squares estimation method (DWLSMV; Muthén, 1984) can be applied to estimate confirmatory factor models for ordinal item responses (e.g., the GRM or GMFRM with a latent variable representation and a probit link function). In this three-stage approach, only the univariate or bivariate frequencies of items (or virtual items, respectively) are used to estimate item thresholds and the polychoric correlations of all items in the first two stages. In the third stage, the item thresholds and the polychoric correlation matrix are estimated as a function of an unknown parameter describing the threshold and covariance structure. DWLSMV estimation can be implemented in the **lavaan** package. In complex rating designs with many raters, not many data are available on virtual items (the response of a particular rater to a particular item) and the estimation of thresholds and polychoric correlations becomes unstable. Therefore, the DWLSMV cannot be reliably applied in these situations.

Composite maximum likelihood estimation (see Varin, Reid, & Firth, 2011, for a review) uses a modified optimization function in such a way that only parts of the data are modeled. We will focus only on the case that specifies a likelihood function for all pairs of items (or virtual items). In contrast to DWLSMV estimation, composite methods are one-stage methods and are applicable to complex rating designs. Composite marginal maximum likelihood estimation (CMML; also labeled as pairwise likelihood estimation) is an estimation method of the confirmatory factor model for ordinal data with a latent variable representation under the probit link function (Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012). The estimation is based on the frequencies of the bivariate cross tables of all item pairs. These frequencies are modeled as functions of the model-implied likelihood function, which can be simply evaluated as a function of the unknown model parameters because it can be computed based on the bivariate normal distribution function. Therefore, the estimation method is computationally efficient and arbitrary missing patterns in rating designs can be easily handled. Many variants of the GMFRM in the GRM formulation can be efficiently estimated. Item discriminations, factor covariances, or residual correlations can be estimated as functions of the basis parameters, like in the LLTM, which makes it possible to test the specific hypotheses of rater effects. The CMML estimation approach is implemented in the **lavaan** package and in the `immer::immer_cmml()` function, with a particular emphasis on LLTM representations of the model parameters. The related approach of Garner and Engelhard (2009) is also based on eliminating person parameters by considering pairwise conditional probabilities. However, they propose that model parameters should be estimated by a noniterative algorithm based on eigenvalues on the incidence matrix of pairwise frequencies (the so-called eigenvector method).

As an alternative to CMML, a composite estimation method based on the CML principle can be employed. Composite conditional maximum likelihood estimation (CCML) evaluates the conditional likelihood for pairs of items. Hence, it is also based on only the bivariate information of the dataset. The CCML approach has been proposed for the LLTM for dichotomous data (Zwinderman, 1995) but it can be generalized to polytomous items; our implementation can be found in the `immer::immer_ccml()` function. To this end, Rasch-MFRMs can be estimated more efficiently with CCML than with CML in complex rating designs.

In recent years, Bayesian estimation approaches such as Markov chain Monte Carlo (MCMC) have become very popular due to the availability of very flexible general purpose Bayesian software programs such as BUGS, JAGS, or Stan. In a nutshell, the MCMC approach is a simulation-based stochastic estimation algorithm, which uses random draws of latent variables (person ability, random effects) and model parameters conditional on the information contained in the dataset. The MCMC approach is often seen as being computationally superior to ML estimation for IRT models with many latent variables (Patz & Junker, 1999). In the GMFRM, the random effect person ability as well as the person-item and person-rater effects are estimated. It is relatively easy to estimate this model in a Gibbs sampling approach (Wang et al., 2014). The **immer** package provides a wrapper function for the JAGS software (Plummer, 2003) in the `immer::immer_gmfrm()` function. The HRM of Patz et al. (2002) is mostly estimated with MCMC methods although ML estimation is also possible (DeCarlo et al., 2011). A Metropolis-Hastings within Gibbs sampling algorithm is employed in the `immer::immer_hrm()` function.

It should be emphasized that MCMC estimates are asymptotically equivalent to ML estimates. Hence, MCMC can also be used in applications without a primary focus on Bayesian statistical inference. In IRT models for raters, informative prior distributions decode prior knowledge about parameters in the Bayesian approach. Rater models are often highly parameterized and researchers aim to avoid statistical overfitting. For example, many item-specific rater effects are estimated in a rater model but only practically relevant effects should be signaled by the model. An informative prior normal distribution with a mean of zero and a variance of 0.01 assumes that most rater effects are small. Only those rater effects with large values are estimated as being significantly different from zero (see Muthén & Asparouhov, 2012, for the application of prior distributions in differential item functioning). In an alternative interpretation, model parameters are regularized in such a way that all nonsignificant effects are reduced to zero, which provides a more focused view on the most relevant effects. Similarly, so-called penalized ML estimation has been proposed as a regularization procedure under the ML paradigm for assessing differential item functioning (Tutz & Schauberger, 2015). In the same manner, rater effects can be regularized in a penalized ML approach of a Rasch-MFRM, which will probably be implemented in the TAM package in the near future.

4 Choosing an appropriate rater model

The question of how to choose a suitable model involves an examination of the assumptions, expectations, and properties of the statistical models. In the following, we try to provide a balanced view of advantages and disadvantages of the rater models presented in Section 2.

It has been argued in Section 2 that typical rating designs imply the existence of local dependence caused by person-item and person-rater interactions. While the GMFRM deals with both sources of dependence, the HRM (either in the Patz et al., 2002 or the HRM-SDT specification) only considers additional dependence caused by person-item interactions. It can be argued that, for analytic ratings, halo effects (person-rater interactions) play only a minor role and that therefore, the HRM often fits empirical datasets sufficiently well. The GMFRM and HRM have the advantage that they typically provide a good model (or are at least superior to the MFRM) and provide adequate reliability estimates of person parameters, as sources of local dependence are explicitly modeled. By applying one of the two model classes, a researcher puts substantial emphasis on local dependence because the meaning of all of the model parameters (item parameters, rater effects, and distribution parameters) is coupled with the modeled dependence. In particular, the item and rater parameters in the GMFRM must be interpreted as being conditional on person ability and random person-item and person-rater effects. If the variances of the random item effects substantially differ from each other, item difficulties can no longer be directly compared to each other because they operate on different metrics. A comparison can be made if the random effects are integrated out to form the conditional item response probabilities (see Section 2.3). In addition, the parallel appearance of person ability and random item and rater effects in the GMFRM implies that there is no unique (weighted) maximum likelihood estimator (WLE, Warm, 1989) for the person parameter. Only the mean of the marginal posterior distribution (i.e., the expected value of the posterior distribution, EAP) can be used as a person ability estimate. It should be noted that even in the case of equal discrimination parameters in the GMFRM with random effects, the sum score is no longer a sufficient statistic of the EAP because the ratings are weighted in such a way that ratings corresponding to random effects with smaller variances receive larger weights, while random effects with larger variances receive lower weights. Such a weighting scheme is not always favored, especially in applications in which the person ability estimate is of vital importance for the person itself (e.g., in feedback or in an examination). The HRM and GMFRM are more computationally demanding than unidimensional rater models and this could be seen as a disadvantage for practitioners. We think that this problem can be solved with sufficient computational resources and is not a real limitation in the application of more complex models.

Admittedly, the HRM and GMFRM can probably also not model aspects of the data in order to describe the complex rating behavior. For example, raters can function differently between persons (e.g., Eckes, 2005) or there could be rater drift during a

rating administration (e.g., Leckie & Baird, 2011). Persons are also often clustered within organizational units (e.g., in universities, classes, courses, groups of peers, etc.). This clustering induces additional dependence, which remains nonmodeled in the HRM or GMFRM. However, these aspects are mostly not of major interest in statistical analysis and will be considered as a nuisance (and therefore ignored in the statistical model). Hence, a misspecified likelihood will almost always be the consequence, and pseudo-likelihood estimation is essentially employed, which requires robust ML standard errors (White, 1982). Model parameters resulting from pseudo-likelihood estimation can be interpreted as estimates of some of the population parameters of an assumed statistical model obtained by repeated sampling processes (of persons, raters, clusters, etc.) with comparable assumptions.

In the Rasch-MFRM, the model parameters can be interpreted as being conditional on person ability. The Rasch-MFRM models rater behavior by using a restricted PCM. It has the advantage of computational simplicity as (bias-corrected) JML estimation is computationally fast. Moreover, the sum score of the item responses of a person is a sufficient statistic for the person parameter (WLE, MLE), which facilitates interpretation because of the equal weighting of all the ratings. Rasch-MFRMs assume local stochastic independence and therefore ignore possible dependencies caused by rating the same items or ratings by the same raters. Interestingly, the assumption of local independence in the application of a unidimensional item response model can essentially be reduced to the assumption that residuals cancel out on average. This means that it is assumed that positive and negative local dependence cancel each other out. This assumption is defensible if person ability is interpreted as a major dimension that is statistically extracted from the dataset. Possible violations caused by local independence are regarded as a nuisance factor in statistical modeling. If the sole argument for applying the HRM or the GMFRM is to obtain correct standard errors or adequate reliability estimates for person parameters, we think that this choice is unfounded and that the Rasch-MFRM should be considered instead. The application of the Rasch-MFRM under the local independence assumption should be contrasted with the GMFRM, in which the appearance of random effects only allows for positive local dependence. The nonmodeled positive dependence in the Rasch-MFRM implies that the reliability of the person parameters is underestimated and, therefore, procedures correcting for local dependence have been proposed (Bock, Brennan, & Muraki, 2002). With respect to model parameters such as item difficulties or rater severity effects, robust ML standard errors should be used because the Rasch-MFRM will typically employ a misspecified likelihood function. Notably, this pseudo-likelihood estimation nevertheless provides consistent parameter estimates under repeated sampling assumptions because, asymptotically, the (Kullback-Leibler) distance between a true complex (and unknown) distribution and an assumed parameterized distribution is minimized (White, 1982). As a consequence, the parameter estimates of the Rasch-MFRM for different samples are only comparable (or can only be linked to each other) if similar rating designs are employed that ensure that the extent of (ignored) local dependence remains similar in different samples. When this condition is fulfilled, the use

of the Rasch-MFRM can be justified in applications if the calculation of standard errors for model parameters and person parameters is modified appropriately. This is the case for numerous simulation studies that have aimed to show that applying the Rasch-MFRM to data generated by a GMFRM provides biased parameter estimates (e.g., Wang et al., 2014) because the two models parameterize item response functions in different ways and therefore preclude any legitimate comparison (see Luecht & Ackerman, 2018, for a general discussion about the generalizability of findings from simulation studies in IRT).

It can be expected that a GMFRM including item or rater discrimination parameters will almost always provide a better fit than the Rasch-MFRM. However, we believe that items and raters should be equally weighted as in the Rasch-MFRM because, in applications, the latent construct of interest is defined by having equal contributions of items and persons (see Reckase, 2017 for such a domain sampling perspective). Otherwise, the psychometric model would reweigh the contributions of items and persons in a completely data-driven way, which could be regarded as a threat to validity (Brennan, 2001b). It is sometimes argued in the literature (e.g., Bond & Fox, 2001) that Rasch models have many desirable statistical properties that are not fulfilled in a GMFRM with discrimination parameters (2PL). Maybe a reason for the existence of several myths about the Rasch model could be the property of so-called specific objectivity (Fischer, 1995), which is only guaranteed by the Rasch model and enables the separation of person and item properties in an additive way. Some researchers incorrectly interpret this property as a sample independence of person and item parameters. However, if a statistical model (Rasch or 2PL) holds under the assumption of invariant item parameters (e.g., the same parameters can be applied for specified subpopulations of persons), unbiased comparisons for arbitrary selections of items are possible for both Rasch and 2PL models. The Rasch model has the distinctive advantage that, due to the existence of the sufficient statistic of the sum score, CML estimation can be conducted. However, CML estimation and MML estimation, usually performed for 2PL models, will both provide consistent parameter estimates. Therefore, some researchers' preference for the Rasch model instead of the 2PL model can statistically only be justified by the feasibility of CML estimation (see van der Linden, 1994, for more detailed arguments), but CML is inferior to MML estimation in finite samples. In summary, we believe that the advantage of using the Rasch-MFRM can only be argued by using validity reasons related to the equal weighting property and to the ease of parameter interpretation; we do not believe that it can be argued that the Rasch-MFRM has superior statistical and measurement-related properties.

The rater models discussed above place person ability and model parameters onto a metric of a latent variable, namely, the logit metric or a probit metric. Sometimes, it is preferable to use the original metric of raw scores for interpretational purposes. This seems to be particularly true for research settings in which people with less training in psychometrics are involved. As it has been argued in Section 2.4, the application of CSMs or G-theory models to ordinal data structurally represents the mean and covariance structure of the data and provides consistent parameter estimates, although the assumed normal distribution is misspecified. An important application is the computation of fair

scores (see Eckes, 2015), which adjust person parameter estimates for systematic rater effects. While the use of fair scores in the logit metric of the Rasch-MFRM entails a bias at the boundary of ratings scales (especially for datasets with only few ratings per person), employing the original metric by using a normal distribution model avoids this bias.

Finally, the role of the fit of particular entities (items, raters, persons) or of the whole model has to be considered. From a strictly psychometric perspective, the application of the model fit of an IRT model from a random persons perspective treats model fit as the discrepancy between an observed and a model-implied covariance structure with respect to the items (or virtual items). Therefore, items are considered as being fixed and nonexchangeable, and a possible replication of the experiment must involve the same items and same raters (Brennan, 2011). The application of the G-theory (or classical test theory, CTT) only makes assumptions about random sampling with respect to persons, items, and raters. As the samples are thought to be representative with respect to corresponding populations, all observations have to be equally weighted in the statistical model. It seems that the adequacy of applying G-theory models with equal discrimination parameters can be tested against the application of models in which different discrimination parameters are allowed. However, the perspective of fit does not play a role in the G-theory as the model is only intended to represent the sampling process. Hence, G-theory models or CTT models essentially require fewer assumptions than IRT models (see Brennan, 2011) and, therefore, they allow for broader generalizations. Unfortunately, this fact is often overlooked in applied research and even in parts of the psychometric literature.

To sum up, we have discussed the possible arguments for choosing one of the classes of models for human ratings. These models have different assumptions, which can often be simultaneously defended for a single dataset under different research perspectives or with different uses of model parameters. Applied researchers should be cautious of the psychometric literature that promotes the superiority of one model class over another and justifies its recommendations mainly based on the results of simulation studies.

5 Empirical application in R

In this section, we illustrate the application of several IRT models on a sample dataset and show how they can be estimated within R. The sample dataset is contained in the **immer** package and has the name `data.ptam4`. It comprises 592 ratings for a single essay written by 209 students and rated by ten raters on three items. 39 students were rated by all ten raters, one student by nine raters, 17 students received ratings from two, three or four raters, and 152 students had only ratings from a single rater. Each row in the sample dataset `data.ptam4` includes all ratings of a rater on all items corresponding to an essay of a student. The structure of the dataset can be inspected in R by using the `head()` function.

It can be seen that the student with identifier (variable `idstud`) 10010 has two rows in the dataset which means that she or he received ratings from two raters (variable `rater`) 844 and 802. Ratings were provided on three items `crit2`, `crit3` and `crit4` on a four-point scale (with integer values 0, 1, 2, and 3).

Complete syntax for the specification of all models in this section is provided by a vignette which is included in the **immer** package.

```
R> data(data.ptam4, package="immer")
R> dat <- data.ptam4
R> head(dat)
```

```
  idstud rater crit2 crit3 crit4
1  10005   802     3     3     2
2  10009   802     2     2     1
3  10010   844     0     1     2
4  10010   802     2     2     1
5  10014   837     1     2     2
6  10014   824     0     2     2
```

Item response models for a single item

Before analyzing the complete rating dataset with three items, we investigate rater effects based on only a single item "crit2". We use a dataset in a so called wide format in which columns refer to ratings of a single rater. In our analysis, we use ratings of 40 students who received multiple ratings from ten raters. Only one student was unintentionally rated by only nine raters. The dataset can be attached as `data.ptam4wide` from the **immer** package.

Table 1:
Descriptive Statistics for Item "crit2"

Rater	Cat0	Cat1	Cat2	Cat3	M	SD	Cor
R802	.10	.38	.38	.15	1.58	0.87	.76
R803	.33	.38	.18	.13	1.10	1.01	.79
R810	.20	.38	.33	.10	1.33	0.92	.86
R816	.25	.25	.35	.15	1.40	1.03	.80
R820	.03	.46	.38	.13	1.62	0.75	.69
R824	.23	.40	.25	.13	1.28	0.96	.78
R831	.18	.33	.35	.15	1.48	0.96	.87
R835	.38	.28	.23	.13	1.10	1.06	.76
R837	.08	.30	.35	.28	1.83	0.93	.79
R844	.30	.25	.25	.20	1.35	1.12	.78

Note: Cat0, Cat1, Cat2, Cat3 = relative frequencies for categories 0, 1, 2, 3; Cor = correlation of rating of a single rater with average score across all raters

Table 1 displays descriptive statistics for the rating dataset for item "crit2". Category frequencies, the mean, the standard deviation and the correlation of a rating with the aver-

age score across all ratings are shown in the table. By comparing the means, it is evident that Raters 803 and 835 are severe while Rater 837 is lenient. Moreover, by inspecting relative frequencies and the standard deviations, Rater 820 shows a centrality tendency while Rater 844 can be characterized by an extremity tendency. Finally, by considering the correlation with the average score, Rater 820 exhibits the lowest agreement, while Raters 810 and 831 show the largest extent of agreement.

In the next step, we apply several item response models to the rating dataset involving the single item "crit2" (see Wolfe, 2014, M. Wu, 2017 for applications of this approach). In this approach, an item refers to a single rater and each rater is parameterized by its own set of parameters. First, it is assumed that a continuous variable is used for modelling the ratings of the four-point scale item. We fit the PCM with an assumption of homogeneous rater effects (i.e., all raters possess the same set of parameters), the PCM, and the GPCM (Models M01, M02, and M03). Second, we follow the principle of the HRM in which true ratings of an item are modelled. Therefore, we specify located latent class Rasch models (LOCLCA; Formann, 1985) which parameterize the response functions of the raters by the PCM but assume a discrete ability variable. The locations of these latent classes on the θ metric and the class probabilities are estimated. In our analysis, we fit LOCLCA with three, four and five latent classes (Models M13, M14, and M15). For the four-point scale item, a LOCLCA with four latent classes would be expected if true ratings can be empirically identified.

Table 2:
Model Comparisons for Item Response Models for Item "crit2"

Label	Model	Deviance	#par	AIC	BIC
M01	PCM equal	861.03	4	869	**876**
M02	PCM	**785.58**	31	**848**	900
M03	GPCM	774.45	40	854	922
M13	LOCLCA(3)	808.14	34	876	934
M14	LOCLCA(4)	775.52	36	848	**908**
M15	LOCLCA(5)	**769.24**	38	**845**	909

Note: #par = number of estimated parameters; PCM equal = partial credit model in which parameters for all raters were constrained to be equal; LOCLCA(k) = Located class analysis with k located latent classes and the PCM is used as the item response function.

Table 2 contains deviances and information criteria for the fitted models. Model selection can be conducted by using differences of deviance values of nested models and performing a likelihood ratio test (LRT) or by considering models with smallest information criteria AIC and BIC. When comparing models M01, M02 and M03, it turned out that the model with equal parameters for raters must be rejected which means that raters differ with respect to their rating behavior. The GPCM did not fit the data significantly better than the PCM although the small sample size ($N = 40$) has to be considered. As an example, we show how to fit the PCM using the **TAM** package and discuss parts of the summary

output (dat2 is the dataset data.ptam4wide).

```
R> items <- c("crit2","crit3","crit4")
R> mod02 <- TAM::tam.mml(  resp=dat2[,items] , irtmodel="PCM2")
R> summary(mod)
```

```
Item Parameters -A*Xsi
   item  N     M xsi.item AXsi_.Cat1 AXsi_.Cat2 AXsi_.Cat3 B.Cat1.Dim1
1  R802 40 1.425    0.274     -2.923     -2.414      0.823           1
2  R803 40 1.300    0.714     -1.033     -0.730      2.142           1
3  R810 40 1.500   -0.086     -2.960     -2.047     -0.257           1
4  R816 40 1.600   -0.401     -3.892     -3.814     -1.203           1
5  R820 39 1.513   -0.011     -4.095     -3.747     -0.033           1
6  R824 40 1.450    0.333     -2.867     -2.745      0.998           1
7  R831 40 1.425    0.180     -2.263     -1.177      0.540           1
8  R835 40 1.450    0.228     -3.922     -3.602      0.683           1
9  R837 40 2.075   -2.073     -5.491     -7.872     -6.219           1
10 R844 40 1.600   -0.199     -2.812     -2.944     -0.597           1
```

The argument irtmodel="PCM2" requests the Andrich (1978) parameterization of the PCM. The column xsi.item contains the item difficulty of the PCM which can be interpreted as rater severity/leniency. The most lenient Rater 837 has the smallest item difficulty (i.e., rater difficulty) while for the most severe Rater 803 the largest parameter was obtained. The columns AXsi_.Cat1, AXsi_.Cat2 and AXsi_.Cat3 include rater-category parameters which assess aspects of severity/leniency or centrality/extremity behavior of the raters. The rater parameters can most easily be interpreted by computing Thurstonian thresholds from the PCM using the TAM::tam.threshold() function.

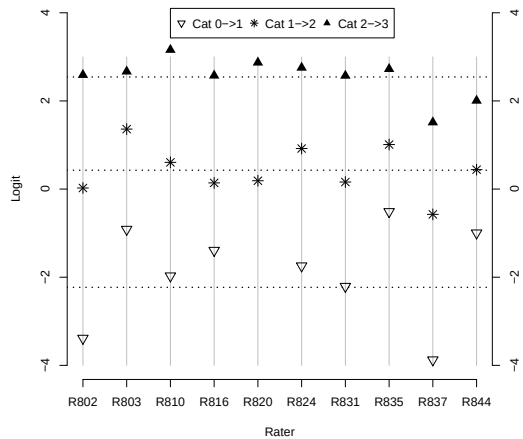

Figure 2:
Rater thresholds from the PCM (Model M02)

In Figure 2, the thresholds for all raters and all categories are depicted. It is evident that

Raters 802 and 837 are lenient with respect to using the zero category while the opposite is true for Rater 835. Summing up, it can be seen that the variability of the thresholds among raters between 0 and 1 is much larger than for the thresholds between 1 and 2 and 2 and 3. This means that the raters show less agreement for rating students in lower categories, but more agreement in rating higher categories.

We also compute infit statistics for raters (Eckes, 2015) based on the PCM (Model M02) using the `TAM::msq.itemfit()` function. Raters 810 and 831 which showed the highest agreement with the average rating (see Table 1) have the lowest infit statistics (.77 and .75, respectively) which can be interpreted as overfit. The largest infit statistics were observed for Raters 835 and 844 (1.14 and 1.19, respectively) which indicates underfit of these two raters. The GPCM can be fitted using the `TAM::tam.mml.2pl()` function using the argument `irtmodel="GPCM"`. It turned out that Raters 810 and 831 have the largest rater discriminations (4.25 and 4.39, respectively).

The model fit for different LOCLCAs are shown in Table 2. It should be emphasized that the LOCLCA with four or five latent classes has a slightly superior fit to the PCM which assumes a continuous ability. Although the LOCLCA with four classes could be preferred because it can be more easily interpreted, we present the results of the LOCLCA with five classes. The LOCLCA can be fitted using the `TAM::tamaan()` function which allows the specification of IRT models similarly to the **lavaan** package.

```
R> tammodel <- "
R+ ANALYSIS:
R+ TYPE=LOCLCA;   # type of the model
R+ NCLASSES(5);   # 5 classes
R+ NSTARTS(10,30); # 10 random starts with 30 iterations
R+ LAVAAN MODEL:
R+   F =~ R802__R844
R+     "
R> mod15 <- TAM::tamaan( tammodel , resp=dat2 )
R> summary(mod)
```

```
Cluster locations
        V1  prob
Cl1 -9.990 0.048
Cl2 -2.814 0.108
Cl3 -0.124 0.334
Cl4  1.463 0.335
Cl5  3.415 0.175
------------------------------------------------------------
Item Response Probabilities
   item itemno Cat Class1 Class2 Class3 Class4 Class5
1  R802      1   0 0.9988 0.3835 0.0284 0.0024 0.0000
2  R802      1   1 0.0012 0.5975 0.6526 0.2641 0.0242
3  R802      1   2 0.0000 0.0190 0.3056 0.6048 0.3903
4  R802      1   3 0.0000 0.0001 0.0133 0.1288 0.5855
```

```
[...]
------------------------------------------------------------
Class-Specific Item Means
   item itemno Class1 Class2 Class3 Class4 Class5
1  R802      1 0.0012 0.6357 1.3038 1.8600 2.5612
2  R803      2 0.0001 0.0747 0.6278 1.3383 2.4807
3  R810      3 0.0002 0.2063 1.0124 1.6340 2.3854
4  R816      4 0.0001 0.1067 0.9765 1.8310 2.5666
5  R820      5 0.0553 1.0133 1.2978 1.7869 2.4802
6  R824      6 0.0001 0.1668 0.9042 1.5541 2.4835
7  R831      7 0.0002 0.2583 1.1573 1.8270 2.5646
8  R835      8 0.0000 0.0455 0.5335 1.4349 2.4938
9  R837      9 0.0027 0.8269 1.5150 2.2048 2.8086
10 R844     10 0.0001 0.0719 0.7754 1.8169 2.7131
```

The latent classes from the model output can be interpreted as latent ratings. By inspecting item response probabilities and class-specific item averages, latent classes 1 and 2 can be associated with "true" category 0, and latent classes 3, 4 and 5 can be associated with "true" categories 1, 2 and 3. Note that Class 1 includes students which were (very probably) rated as 0 by all raters while raters differed in their ratings for students in Class 2. Raters 802, 820 and 837 rated a substantial portion of students in Class 2 into categories 1, 2 or 3 while all other raters mostly rated students into category 0. Moreover, from the output it can be also concluded that Rater 837 is the most lenient one.

G-theory models

In the following analyses, we use the full datasets including three items, ten raters and 209 students. As a preliminary analysis to more complex item response models, we fit G-theory models (specified as linear mixed effects models) for assessing the amount of variance which can be attributed to different sources. We estimate G-theory models using the **lme4** package. In order to achieve this, the dataset has to be converted into a long format in which one row refers to the combination of a student, a rater and an item. The needed structure has already been prepared as the dataset `data.ptam4long` in the **immer** package. Four different G-theory models are fitted (Models M21, M22, M23 and M24). The first three models assume homogeneous variance components (for random effects of items or raters) while the last model allows for item-specific or rater-specific variances of random effects. The G-theory Model M23 including person, person-item and person-rater random effects can be estimated using the following syntax (`value` denotes the variables which include all ratings for students, items and raters)

```
R> mod23 <- lme4::lmer( value ~ rater*item + ( 1 | idstud ) +
R+      ( 1 | idstud:item ) + ( 1 | idstud:rater), data = data.ptam4long )
R> summary(mod23)
```

```
Random effects:
 Groups       Name        Variance Std.Dev.
 idstud:item  (Intercept) 0.06497  0.2549
 idstud:rater (Intercept) 0.09344  0.3057
 idstud       (Intercept) 0.28119  0.5303
 Residual                 0.21512  0.4638
Number of obs: 1776, groups: idstud:item, 627; idstud:rater, 592; idstud, 209
```

In this model, rater-specific item means are allowed (fixed effects `item*rater`). It can be seen from the output that a large part of the variance corresponds to student ability. Interestingly, the variance component due to person-rater interactions (i.e., halo effects) is slightly larger than the amount of dependence due to person-item interactions. This finding sheds some light on the application of HRM which only handles dependency due to random item effects but not to random rater effects.

Table 3:
Variance Component Estimates from G-theory Models

Variance	Model M21	Model M22	Model M23
p	.331	.323	.281
$p \times i$	—	.044	.065
$p \times r$	—	—	.093
Residual	.334	.299	.215

Note: p = persons; i = items; r = raters

In Table 3, the variance component estimates for the first three models are shown. When comparing Model M21 and M22, it can be seen that most part of the variance of the item effect ($p \times i$) is confounded with the residual variance in Model M21. When including the random rater effect ($p \times r$) in Model 23, a substantial part of the true score variance is captured which shows that neglecting dependency due to halo effects results in overly optimistic reliability estimates because the true score variance is overestimated.

Finally, we show how to estimate a G-theory model with heterogeneous variance components (Model M24). The specification is a bit cumbersome when done manually because dummy variables for all items (e.g., `I_crit2`) and all raters (e.g., `R_802`) are involved in the model specification.

```
R> mod24 <- lme4::lmer( value ~ rater * item + (1 | idstud) +
R+    (0 + I_crit4 | idstud:item) +  (0 + I_crit3 | idstud:item) +
R+    (0 + I_crit2 | idstud:item) +  (0 + R_844 | idstud:rater) +
R+    (0 + R_837 | idstud:rater) +  (0 + R_835 | idstud:rater) +
R+    (0 + R_831 | idstud:rater) +  (0 + R_824 | idstud:rater) +
R+    (0 + R_820 | idstud:rater) + (0 + R_816 | idstud:rater) +
R+    (0 + R_810 | idstud:rater) +  (0 + R_803 | idstud:rater) +
R+    (0 + R_802 | idstud:rater), data= data.ptam4long)
```

The variance component estimates of person-item interactions (`idstud:item`) were estimated as .094 (Item “crit2”), .000 (Item “crit3”) and .092 (Item “crit4”) showing that

no local dependency was introduced for "crit3". The variance component estimates of person-rater interactions (`idstud:rater`) showed a considerable amount of variation (M= .096, SD=.059, Min=.027, Max=.204).

Many-facet rater models

In this subsection, we illustrate the application of several MFRMs. In a first series of models, we fit Rasch-MFRMs which assume equal item and rater discrimination parameters (Models M31, ..., M36). In a second series of models, we fit MFRMs which allow the inclusion of item and rater discrimination parameters (Models M41, ..., M46).

In a Rasch-MFRM, the item response function for person p, item i, rater r and category k is given as $P(X_{pir} = k|\theta_p) \propto \exp(k\theta_p - b_{irk})$. Different constrained versions for parameters b_{irk} can be estimated. These versions can be defined using design matrices or – more conveniently – using the formula language in R when fitting Rasch-MFRMs with the `TAM::tam.mml.mfr()` function in the **TAM** package. For example, the formula `~ item*step + rater` for facets items, raters and steps (i.e., categories) corresponds to the constraint $b_{irk} = b_{ik} + b_r$. In principle, formulas of arbitrary complexity and an arbitrary number of facets can be specified in the **TAM** package using the argument `formulaA`. The Rasch-MFRM `~ item*step + rater` (Model M32) can be estimated using the following syntax (`dat` is the dataset `data.ptam4`)

```
R> facets <- dat[, "rater", drop=FALSE ]
R> mod32 <- TAM::tam.mml.mfr( dat[,items], facets=facets,
R+             formulaA = ~ item*step + rater, pid=dat$pid )
R> summary(mod32)
```

```
Item Facet Parameters Xsi
[...]
7     rater802     rater -0.118  0.101
8     rater803     rater  1.247  0.101
9     rater810     rater -0.052  0.099
10    rater816     rater -0.017  0.101
11    rater820     rater -0.412  0.099
12    rater824     rater  0.024  0.099
13    rater831     rater  0.169  0.100
14    rater835     rater  0.666  0.100
15    rater837     rater -1.483  0.099
16    rater844     rater -0.023  0.300
[...]
```

The function automatically creates virtual items for estimating the constrained PCM. The estimated item and rater parameters can be found in output section `Item Facet Parameters Xsi` (only rater parameters are displayed). The main rater effects in this section indicate the extent of leniency/severity tendencies. These rater effects are almost perfectly correlated with the means for each rater (across all items) which can be expected because these means are sufficient statistics for the rater effects.

Table 4:
Model Comparisons of Different Rasch-MFRMs

Label	formulaA	Deviance	#par	AIC	BIC
M31	~ item*step	3802.42	10	3822	**3866**
M32	~ item*step+rater	3763.23	19	3801	3885
M33	~ item*step+rater*step	3693.46	37	3767	3930
M34	~ item*step+rater*item	3699.89	37	3774	3936
M35	~ item*step+rater*item+rater*step	3632.30	55	**3742**	3983
M36	~ item*rater*step	**3562.38**	91	3744	4143

Note: formulaA = formula specification of Rasch-MFRM; #par = number of estimated parameters.

In Table 4, model comparisons of different Rasch-MFRMs are shown. It can be seen that model fit improves when interaction effects of raters and items or raters and categories are included. The most complex model which assumes a PCM for all virtual items based on combinations of items and raters would be favored based on a LRT but not based on information criteria. It should be noted that information criteria are not to be expected to provide valid statistical inference in case of incomplete designs[1]. This finding highlights that the specification of rater models should not stop with modelling severity/leniency tendencies as other rater effects can be of similar or larger importance.

In a second series of models, we investigate whether differences in discriminations of items or raters can be found. Based on the item response function $P(X_{pir} = k|\theta_p) \propto \exp(ka_{ir}\theta_p - b_{irk})$, different specifications for the discrimination parameter a_{ir} are employed. These models can be estimated using the `sirt::rm.facets()` function. Different specifications for the discrimination parameters can be requested by using the arguments `est.a.item` and `est.a.rater`. The following syntax shows how to estimate Model M44 (see also Table 5) which includes item and rater discriminations in a multiplicative way (i.e., $a_{ir} = a_i a_r$). A model based on virtual items in which all PCM (or GPCM) parameters are estimated can be requested by the argument `rater_item_int=TRUE`.

```
R> mod44 <- sirt::rm.facets( dat[ , items], rater=dat$rater, pid=dat$pid,
R+           est.a.item=TRUE, est.a.rater=TRUE, reference_rater="831",
R+           rater_item_int=FALSE)
R> summary(mod44)

Item Parameters
   item   N     M tau.Cat1 tau.Cat2 tau.Cat3     a  delta delta_cent
1 crit2 592 1.409   -2.244   -2.053    0.542 0.889  0.181      0.368
```

[1] A large fraction of students in the dataset only received a single rating. With three items on a four-point scale, 10 parameters can be estimated for these students (9 item parameters and 1 variance parameter). However, in the Rasch-MFRM specification all students are penalized in the AIC formula by the total number of parameters which refers to a response pattern for students which have received multiple markings from all raters (90 item parameters, 1 variance parameter). Therefore, the number of estimated parameters in the AIC formula must be an overestimate of penalization in incomplete designs.

```
2 crit3 592 1.586   -5.166   -5.702   -1.675 1.475 -0.558     -0.371
3 crit4 592 1.508   -3.342   -3.299   -0.555 0.762 -0.185      0.003
-----------------------------------------------------------------------
Rater Parameters
   rater   N     M      b     a thresh b.cent a.cent
1    802 174 1.540  0.147 0.989  0.146  0.095  1.053
2    803 183 1.158  0.959 1.263  1.211  0.906  1.327
3    810 183 1.508 -0.306 0.972 -0.298 -0.358  1.036
4    816 171 1.503  0.264 0.972  0.257  0.212  1.035
5    820 180 1.606 -0.544 0.862 -0.469 -0.597  0.926
6    824 189 1.492  0.129 1.093  0.141  0.077  1.157
7    831 177 1.446  0.000 1.000  0.000 -0.052  1.064
8    835 171 1.298  0.782 0.605  0.473  0.730  0.669
9    837 180 1.944 -1.049 0.626 -0.656 -1.101  0.690
10   844 168 1.512  0.140 0.980  0.137  0.088  1.044
```

From the output, we see that item "crit3" is more discriminative than the other two items (see column `a`). Further, Raters 803 and 824 are most discriminative (i.e., accurate) while Raters 835 and 837 are least discriminative (i.e., inaccurate) (see column `a.cent`). We also calculated rater infit statistics from the Rasch-MFRM Model M32 (`~ item*step + rater`) and compared these with rater discriminations from Model M42 ($b_{irk} = b_{ik} + b_r$, $a_{ir} = a_r$). Lower rater discriminations tended to result in higher rater infit values ($r = -.33$). It turned out that the relationship of both statistics was stronger ($r = -.59$) when an outlying observation (Rater 803) was removed from the calculation.

Table 5:
Model Comparisons of Different MFRMs

Label	b_{irk}	a_{ir}	Deviance	#par	AIC	BIC
M41	$b_{ik} + b_r$	1	3763.23	18	3799	**3878**
M42	$b_{ik} + b_r$	a_r	3738.55	26	3791	3905
M43	$b_{ik} + b_r$	a_i	3750.96	20	3791	3879
M44	$b_{ik} + b_r$	$a_i a_r$	3724.43	29	3782	3910
M45	b_{irk}	1	3699.89	37	3774	3936
M46	b_{irk}	a_{ir}	**3609.85**	64	**3738**	4018

Note: b_{ir} = specification of item-specific rater intercept; a_{ir} = specification of item-specific rater discrimination #par = number of estimated parameters.

Finally, the model comparison from Table 5 indicated that the most flexible model parameterizing all virtual items by the GPCM would be preferred based on the LRT and AIC. In summary, it can be concluded that Rasch-MFRMs allowing for interaction effects of raters with item or categories or MFRMs with rater discriminations should be preferred from the perspective of model fit to a Rasch-MFRM in which only a main rater severity effect is modelled. We expect that this conclusion will not change if measures of approximate model fit would be employed. This modelling exercise illustrates our argument that a preference of a simpler Rasch-MFRM can (in most applications) only be

justified for validity reasons, that is, when the differential weighting of items and raters by a psychometric model is not warranted.

Generalized many-facet rater models

By employing G-theory models it was observed that there is a substantial amount of variance attributed to person-item and person-rater interactions. Now, a series of GFRMs is fitted in which we allowed item discriminations and we included particular variance components. Four models were specified. Model M51 only contains the random person effect. Model M52 additionally includes the random item effect while in Model M53 the random rater effect is included. Finally, we include all three variance components in Model M54. The **immer** package provides a wrapper function `immer::immer_gmfrm()` to the JAGS software (Plummer, 2003). Model M54 can be estimated using the following syntax.

```
R> mod54 <- immer::immer_gmrfm(dat[,items], rater=dat$rater, pid=dat$idstud,
R+              fe_r="r", re_pi=TRUE, re_pr=TRUE, iter=iter, burnin=burnin)
```

The argument `fe_r` specifies the fixed effects structure of intercepts b_{irk} of the IRT model with respect to raters. Options are `"n"` ($b_{irk} = b_{ik}$), `"r"` ($b_{irk} = b_{ik} + b_r$), `"ir"` ($b_{irk} = b_{ik} + b_{ir}$), `"rk"` ($b_{irk} = b_{ik} + b_{rk}$) or `"a"` (all effects are specified, i.e. all b_{irk} are estimated without constraints). The arguments `re_pi` and `re_pr` indicate whether random effects should be included in the GMFRM. For example, Model M52 can be estimated using `re_pi=TRUE` and `re_pr=FALSE`. We use 50,000 iterations (argument `iter`) and 10,000 burn-in iterations (argument `burnin`) which provided a good convergence behavior of the MCMC estimation approach in our example.

We only briefly discuss the results of Model M54. The variance component estimates varied considerably among items ("crit2": .12, "crit3": .05; "crit4": .66). The rater severities correlated highly with the average rater scores ($r = -.99$) and did also show some variation among raters (SD=.48, Min=-1.06 [Rater 837], Max=0.84 [Rater 803]). The variance estimates for person-rater interactions also exhibited some variability among raters (M=.27, SD=.32). Two Raters 844 (.53) and 803 (1.09) had remarkably high variance estimates indicating that halo effects were strongly present for these raters. Finally, the correlation of the rater variances from the GMFRM (Model M54) and from the G-theory model (Model M24) was .84 indicating that findings remain relatively stable irrespectively of whether the logit or the original metric is chosen.

Hierarchical rater model based on signal detection theory

In the GMFRM, local dependence is taken into account by including additional random effects. In the HRM, ratings are modelled by a hierarchical approach which first assumes that manifest ratings are modelled conditionally on true discrete ratings (signal detection model, SDM). Second, true ratings are modelled by item response functions (item response model, IRM). For both models different specifications can be chosen.

Table 6:
Model Comparisons of Different HRM-SDT Models

Label	IRM	SDM	Deviance	#par	AIC	BIC
M61	PCM	`n`	3540.53	10	3561	3594
M62	PCM	`e`	3530.70	14	3559	3605
M63	PCM	`r`	3314.19	50	3414	3581
M64	PCM	`a`	**3135.89**	130	**3396**	3830
M71	GPCM	`n`	3525.08	12	3549	3589
M72	GPCM	`e`	3512.23	16	3544	3598
M73	GPCM	`r`	3298.47	52	3402	**3576**
M74	GPCM	`a`	3135.18	132	3399	3840

Note: IRM = specified item response model; SDM = specified signal detection model (`n` = no effects; `e` = exchangeable effects for items and raters; `r` = rater effects; `a` = all effects); #par = number of estimated parameters.

In Table 6, different specifications of our fitted models are shown. The IRM uses either the PCM or the GPCM. In the SDM, discrimination parameters d_{ir} and intercept parameters c_{irk} are estimated with several constraints. Regarding our sample dataset it turned out the PCM with a SDM, in which all rater parameters were allowed to be item-specific, showed the best fit (Model M64) in terms of the LRT and AIC.

For facilitating the interpretation, we focus on the discussion of results of Model M63 in which rater effects in the SDM are assumed to be independent of items. The HRM-SDT can be estimated using the `sirt::rm.sdt()` function. To choose the GPCM instead of the PCM one has to use the argument `est.a.item=TRUE`. Different specifications of the SDM can be chosen by using the arguments `est.c.rater` and `est.d.rater`. The estimation of Model M63 can be conducted using the following syntax.

```
R> mod63 <- sirt::rm.sdt( dat[,items], rater=dat$rater, pid=dat$idstud,
R+                est.c.rater="r" , est.d.rater="r")
R> summary(mod63)
```

```
Rater Parameters
   item.rater  N     M     d    c_1    c_2    c_3 c_1.trans c_2.trans c_3.trans
1   crit2-802 58 1.655 4.939  2.590  7.356 13.000     0.524     1.489     2.632
2   crit2-803 61 1.000 3.564  3.264  6.525  9.801     0.916     1.831     2.750
3   crit2-810 61 1.344 4.529  1.213  6.819 11.463     0.268     1.506     2.531
4   crit2-816 57 1.526 5.553  3.336  7.897 14.463     0.601     1.422     2.605
5   crit2-820 60 1.650 7.967 -1.182 11.471 21.195    -0.148     1.440     2.661
6   crit2-824 63 1.365 7.279  4.593 10.668 19.034     0.631     1.466     2.615
7   crit2-831 59 1.373 4.509  1.449  7.464 11.728     0.321     1.655     2.601
8   crit2-835 57 1.035 2.858  1.526  4.762  8.564     0.534     1.666     2.996
9   crit2-837 60 1.833 2.857 -1.314  2.454  6.524    -0.460     0.859     2.283
10  crit2-844 56 1.304 3.679  2.004  4.851  9.323     0.545     1.319     2.534
```

We focus on the interpretation of rater parameters. Raters 820 and 824 are most reliable

because high discrimination parameters d_r were estimated for them. Further, Raters 835 and 837 are least reliable. Severity/leniency and centrality/extremity tendencies can be identified by the intercept parameters c_{rk}. The relative criteria locations $c^*_{rk} = c_{rk}/d_r$ (displayed as `c_1.trans`, `c_2.trans`, `c_3.trans` in the output) indicate the relative "difficulty" for every category of a rater. For raters which do not produce systematic biases, relative criteria locations of .5, 1.5, and 2.5 would be expected for four-point scale items. In Figure 3, these locations are displayed for all raters and all criteria. It can be seen that Raters 820 and 837 are lenient with respect to rating students into the zero category. Rater 803 is more severe because she or he more frequently rates students into the zero category. The standard deviation among raters of relative criteria locations can be computed to assess the uncertainty of rating particular categories. Differentiating students between 0 and 1 showed most variability (SD=.40), while the SD for categories 1 and 2 (SD=.26) and 2 and 3 (SD=.18) was lower. It should be emphasized that a measure of rater severity can be calculated from HRM-SDT output by averaging criteria locations, i.e. computing $\overline{c}^*_r = \sum_k c^*_{rk}/K$.

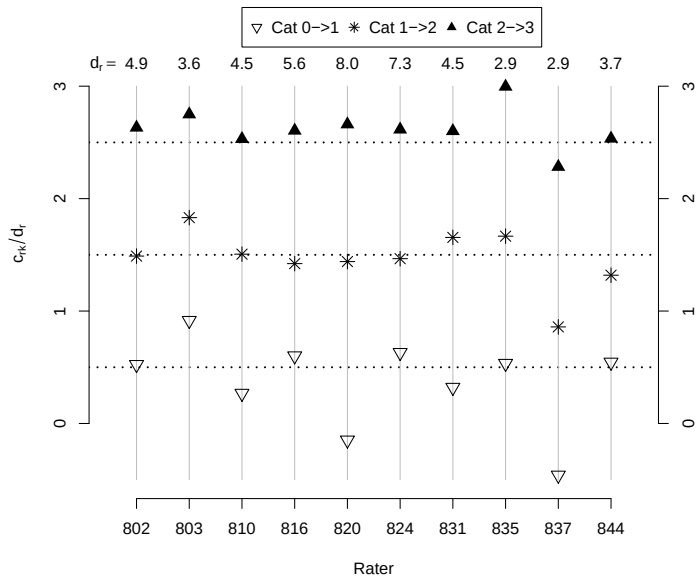

Figure 3: Plots of the relative criteria locations $c^*_{rk} = c_{rk}/d_r$ for the HRM-SDT (Model M63). The solid horizontal lines show intersection points for the underlying distributions.

Hierarchical rater model of Patz et al. (2002)

Finally, we want to fit the alternative HRM of Patz et al. (2002). This model includes rater severity (rater bias) ϕ_{ir} and rater variance ψ_{ir} as rater parameters. Two models are fitted. First, Model M81 assumes that the rater parameters are item independent while in Model

M82 these parameters are specified to vary across items. Different specifications can be chosen by using the arguments `est.phi` and `est.psi` in the `immer::immer_hrm` function. Both models employ the PCM as the IRM. Model M81 can be estimated using the following syntax based on 500,000 iterations and 200,000 burn-in iterations[2].

```
R> mod81 <- immer::immer_hrm( dat[,items], pid=dat$idstud, rater=dat$rater,
R+             est.phi="r", est.psi="r", iter=iter,  burnin=burnin)
R> summary(mod81)

Rater Parameters
    item rater rid N_Rat     M    phi   psi
1  crit2   802   1    58 1.655  0.107 0.383
2  crit2   803   2    61 1.000 -0.303 0.644
3  crit2   810   3    61 1.344  0.156 0.334
4  crit2   816   4    57 1.526  0.078 0.451
5  crit2   820   5    60 1.650  0.208 0.212
6  crit2   824   6    63 1.365  0.035 0.321
7  crit2   831   7    59 1.373  0.057 0.387
8  crit2   835   8    57 1.035 -0.130 0.725
9  crit2   837   9    60 1.833  0.629 0.522
10 crit2   844  10    56 1.304  0.135 0.635
```

The SD of student ability was estimated as 9.42 and was surprisingly high. In the HRM-SDT, a much lower SD of 3.05 was obtained in Model M63. It can be seen in the output of Model M81 that Rater 837 is most lenient because she or he has the highest ϕ value while Rater 803 is most severe. Rater 820 gives the most accurate ratings because she or he has the lowest rater variability ψ while Rater 835 is the least accurate.

The results of the HRM of Patz et al. (2002) (Model M81) should now be compared with the HRM-SDT (Model M63). The correlation of rater precision (i.e., $1/\psi_r$) and rater discrimination (d_r) was relatively high ($r = .88$) indicating that both models reach similar conclusions. Moreover, we correlated the rater severity ϕ_r with the average relative criteria location $\overline{c}_r^*$ of the HRM-SDT. We obtained an almost perfect correlation of $r = -.99$. Therefore, the HRM-SDT also proves useful in assessing rater severity.

Finally, we briefly discuss interesting findings of Model M82. In this HRM, rater severity and rater variances are item-specific. One could question whether all item-rater interaction effects need to be specified. To this end, a F-test based on the MCMC output can be conducted for testing the hypothesis of equal rater severity among items ($\phi_{1r} = \phi_{2r} = \phi_{3r}$) and of equal rater variance ($\psi_{1r} = \psi_{2r} = \psi_{3r}$). This F-test can be computed using the `sirt::mcmc_WaldTest()` function. Seven out of ten raters showed significant differences in item-specific rater severities while for no rater the F-test of the equality of rater variances was significant.

[2] Although much more iterations than in the estimation of the GMFRM were chosen, computation time did not substantially increase because the MCMC algorithm in **immer** is implemented in R using the **Rcpp** package for some parts of the computation.

6 Discussion

In the past sections, we gave an overview of opportunities in psychometric modeling in the field of rater studies and we provided some insight into popular estimation methods. We have introduced models ranging from G-theory, Rasch-MFRM to more recent developments such as hierarchical modeling approaches (GMFRM or HRM). Several basic considerations of assumptions, expectations and properties of the models which are all associated with model choice have been elaborated in Section 4. To take dependencies into account, either between persons and items or persons and raters, the HRM (in the first case) or the GMFRM (for both cases) might be considered. As stated, a drawback might be that the person ability has to be interpreted as conditioned to the modeled dependence. Sometimes it might be more appropriate to treat those dependencies as nuisance factors, in particular, when using the sum scores and the equal weighting of items and raters is favored. In this case, the Rasch-MFRM or G-theory models might be appropriate choices.

To gain an impression which psychometric models are applied in the field of language testing, we conducted a rough literature study. For this study we have used two journals "Language Testing" and "Language Assessment Quarterly". All contributions available online between 2007 and 2017, which have dealt with rater studies, were taken into account. The applied methods were classified into three groups, the Rasch-MFRM, G-theory and a remaining third category "other". The latter category includes both qualitative and quantitative analysis like descriptive statistics, as well as more complex models, such as structural equation models, generalized linear models, etc. It appeared that the Rasch-MFRM is currently the favored model for rater studies within these two selected journals. Over the last 10 years, the Rasch-MFRM has gained popularity. Between 2007 and 2017 the Rasch-MFRM was used in 51.5%, the G-theory in 19.1%, and the "other" methods in 29.4% of the cases. Although this study is not representative for applied methods within the field of language testing, it becomes apparent that there is a considerable preference for the Rasch-MFRM. Similarly, McNamara and Knoch (2012) reviewed the usage of IRT model in the field of language testing between 1984 and 2002 and found that the Rasch model was dominantly used. The authors concluded that development in psychometric methods creates many opportunities, but is also related to challenges of its application by language testers because the interpretation of more complex models is involved.

We think that in the near future more advanced methodological developments will be applied because of a wider availability of software and an increasing familiarity of researchers with recent software. We hope that this paper as well as the fast growing community of package development in R will contribute to reduce the still existing gap between the methodological developments on the one hand and the variety of methods in the field of language testing on the other hand.

Some aspects have yet not been addressed in this paper, but may be also influencing factors for model selection in the broadest sense. First of all, the choice of a rating design

should be emphasized. When choosing a rating design, there are a few possibilities which reach from complete designs in which every rater judges every item to more sophisticated incomplete designs which usually requires a kind of linking. Complete rating designs are often not applicable for economic reasons that is why incomplete designs are chosen. Several types of incomplete rating designs are possible; one of the more established ones might be the practice of using common ratings. Here, a representative selection of persons is rated by all raters, while the remaining ones are rated by one or more, but not by all raters. Another possibility of rating designs which might reduce the venture of choosing not representative common persons are overlap (or incomplete) designs (e.g. DeCarlo, 2010). Different types of overlap designs are feasible. All of them have in common that decisions are required, which persons are allocated to rater, how many persons are rated per rater, and how raters are linked among each other. These considerations or decisions can themselves lead to additional effects (Casabianca & Wolfe, 2017).

Another yet not mentioned aspect of rater models is the development of automated scoring systems, especially in the area of large-scale assessments. From an economical point of view, this approach might be more efficient than human ratings although the validity of automated scoring approaches can be questioned. It should be noted that the discussion about choosing appropriate rater models for human raters is also important for calibrating automated scoring systems because the calibration relies on human ratings (Wind, Wolfe, Engelhard, Foltz, & Rosenstein, 2018).

Acknowledgment

We would like thank Thomas Eckes for stimulating us to contribute to this special issue, and in particular for his consideration and very helpful feedback on previous versions of the manuscript.

References

Aitkin, M. (2016). Expectation maximization algorithm and extensions. In W. van der Linden (Ed.), *Handbook of item response theory. Volume Two: Statistical Tools* (pp. 217–236). Boca Raton: CRC Press.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland Publishing Co.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Arminger, G., & Schoenberg, R. J. (1989). Pseudo maximum likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika*, *54*(3), 409–425.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement*, *34*(8), 607–619.

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., & Zhao, L. (2014). Misspecified mean function regression: making good use of regression models that are wrong. *Sociological Methods & Research*, *43*(3), 422–451.

Bertoli-Barsotti, L., Lando, T., & Punzo, A. (2014). Estimating a Rasch model via fuzzy empirical probability functions. In D. Vicari, A. Okada, G. Ragozini, & C. Weihs (Eds.), *Analysis and Modeling of Complex Data in Behavioral and Social Sciences* (pp. 29–36). Cham, Switzerland: Springer.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26*(4), 364–375.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah: Lawrence Erlbaum.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Spinger.

Brennan, R. L. (2001b). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice*, *20*(4), 6–18.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21.

Casabianca, J. M., & Wolfe, E. W. (2017). The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model. *Psychological Test and Assessment Modeling*, *59*(4), 471–492.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, *42*(1), 53–76.

DeCarlo, L. T. (2010). *Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report RR-10-08). Princeton: ETS.

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333–356.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments (2nd ed.)*. Frankfurt am Main: Peter Lang.

Eckes, T. (2017). Guest Editorial Rater effects: Advances in item response modeling of human ratings–Part I. *Psychological Test and Assessment Modeling*, *59*(4), 443–452.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171–191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. foundations, recent developments, and applications* (pp. 15–31). New York: Springer.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275–299.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 87–111.

Garner, M., & Engelhard, G. (2009). Using paired comparison matrices to estimate parameters of the partial credit Rasch measurement model for rater-mediated assessments. *Journal of Applied Measurement*, *10*(1), 30–41.

Hahn, J., & Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, *72*(4), 1295–1319.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*(4), 577–601.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 395–416.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399–418.

Linacre, J. (1989). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J. (2017). *A user's guide to FACETS: Rasch-model computer programs.* Chicago: Winsteps.com.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, New Jersey: Erlbaum.

Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and*

Practice, Advance online publication. doi: 10.1111/emip.12185

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McLaughlin, J. E., Singer, D., & Cox, W. C. (2017). Candidate evaluation using targeted construct assessment in the multiple mini-interview: A multifaceted Rasch model analysis. *Teaching and Learning in Medicine*, *29*(1), 68–74.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(1), 159–176.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342–366.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Vienna,Austria: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22.

R Core Team. (2018). *R: A language and environment for statistical computing.* Austria: R foundation for statistical computing Vienna.

Reckase, M. D. (2017). *A tale of two models: Sources of confusion in achievement testing*. (Research Report No. RR-17-44). Princeton, NJ: Educational Testing Service.

Robitzsch, A. (2018a). *LAM: Some latent variable models*. R package version 0.2. https://CRAN.R-project.org/package=LAM.

Robitzsch, A. (2018b). *sirt: Supplementary item response theory models*. R package version 2.5. https://CRAN.R-project.org/package=sirt.

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*. R package version 2.9.

https://CRAN.R-project.org/package=TAM.

Robitzsch, A., & Steinfeld, J. (2018). *immer: Item response models for multiple ratings*. R package version 1.0. https://CRAN.R-project.org/package=immer.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rost, J., & Langeheine, R. (1997). A guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 13–37). New York: Waxmann.

Rusch, T., Mair, P., & Hatzinger, R. (2013). *Psychometrics with R: A review of CRAN packages for item response theory*. WU Vienna University of Economics and Business.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores*. Psychometrika Monograph No. 17. Richmond, VA: Psychometric Society.

Tor, E., & Steketee, C. (2011). Rasch analysis on OSCE Data: An illustrative example. *The Australasian Medical Journal*, *4*(6), 339.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 289–316). New York: Springer.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43.

van der Linden, W. J. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice Vol. 2* (pp. 3–24). Norwood, NJ: Ablex Publishing Cooperation.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 5–42.

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). New York: Springer.

Wang, W.-C., Su, C.-M., & Qiu, X.-L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement*, *51*(3), 260–280.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, *50*, 1–25.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192.

Wind, S. A., Wolfe, E. W., Engelhard, G. J., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, *18*(1), 27–49.

Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. White Paper. Pearson Research Reports, Pearson.

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, *79*(4), 453–470.

Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development*, *35*(2), 380–394.

Yuan, K.-H., & Schuster, C. (2013). Overview of statistical estimation methods. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (Vol. 1, pp. 361–387). Oxford: Oxford University Press.

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, *19*(4), 369–375.