The Hierarchical Rater Thresholds Model for Multiple Raters and Multiple Items

Molenaar, D.; Uluman, M.; Tavşancıl, E.; De Boeck, P.

DOI
10.1515/edu-2020-0105
Publication date
2021
Document Version
Final published version
Published in
Open Education Studies
License
CC BY

Research Article

Dylan Molenaar*, Müge Uluman, Ezel Tavşancıl, Paul De Boeck

The Hierarchical Rater Thresholds Model for Multiple Raters and Multiple Items

https://doi.org/10.1515/edu-2020-0105
received April 28, 2020; accepted October 5, 2020.

Abstract: In educational measurement, various methods have been proposed to infer student proficiency from the ratings of multiple items (e.g., essays) by multiple raters. However, suitable models quickly become numerically demanding or even unfeasible as separate latent variables are needed to account for local dependencies between the ratings of the same response. Therefore, in the present paper we derive a flexible approach based on Thurstone's law of categorical judgment. The advantage of this approach is that it can be fit using weighted least squares estimation which is computationally less demanding as compared to most of the previous approaches in the case of an increasing number of latent variables. In addition, the new approach can be applied using existing latent variable modeling software. We illustrate the model on a real dataset from the Trends in International Mathematics and Science Study (TIMMSS) comprising ratings of 10 items by 4 raters for 150 subjects. In addition, we compare the new model to existing models including the facet model, the hierarchical rater model, and the hierarchical rater latent class model.

Keywords: Rating data; Item response theory; Local independence; Hierarchical rater model.

1 Introduction

In the field of educational measurement inferences about students' latent proficiencies underlying educational tests are commonly based on item response theory (IRT) modeling tools. In standard cases, the educational test is purported to measures a single proficiency (unidimensionality) without residual associations between the students' observed scores on the different items (local independence). In educational practice however, items may be clustered because they measure multiple latent variables (e.g., in worded arithmetic problems) or because they concern different testlets. In addition, students may be clustered within classrooms and schools. Neglecting these properties of the testing situation violates the assumptions of the IRT model which can result in biased or inefficient inferences concerning the students' proficiency (Sireci, Wainer & Thissen, 1991; Wainer, 1995; Wainer & Thissen, 1996). Therefore, more sophisticated models need to be considered including multidimensional IRT models (Béguin & Glas, 2001; Reckase, 2009), models for testlets (Bradlow, Wainer, & Wang, 1999; Cai, 2010a), and multilevel or hierarchical IRT models (Adams, Wilson, & Wu, 1997; Mislevy & Bock, 1989; Fox & Glas, 2001).

A specific testing situation in education measurement concerns the case in which the scoring of the student's performance is not objectively possible using an answer key. For instance, in an essay assignment, the grade that a student obtains may depend on the rater that graded the essay. A straightforward and standard solution is to use multiple raters for the same essay. However, the data will violate both the assumption of unidimensionality and local independence, as the items and the students are clustered within the raters. The challenge is to account for these violations by modeling differences among raters in their rater characteristics. The so-called facet model (Linacre, 1989) has previously being used to address this challenge (e.g., Engelhard, 1994, 1996; Wilson & Wang, 1995) and can be considered a common method to model students' ratings. The facet model extends the standard IRT modeling approach by adding a fixed effect for the 'rater severity' to account for differences in rater characteristics. Rater severity refers to the tendency of a rater to assign lower scores to a given response as compared to other raters. That is, the rater severity parameter can be seen as the rater counterpart of the item difficulty parameter.

***Corresponding author: Dylan Molenaar,** Department of Psychology, University of Amsterdam, The Netherlands, E-mail: D.Molenaar@uva.nl
Müge Uluman, Ezel Tavşancıl, Department of Educational Measurement and Evaluation, Ankara University, Turkey
Paul De Boeck, Department of Psychology, Ohio State University, USA

However, although the rater severity parameter accounts for differences across raters, it does not account for the dependency between ratings of the same responses. That is, all rater-item combinations are treated as locally independent. Neglecting the fact that only a subset of these data are locally independent indicators for the students' proficiency has been shown to overestimate the reliability of the students' assessment (see Mariano, 2002, and Wilson & Hoskens, 2001).

As will be discussed below, effort has been devoted to develop suitable models that take both the common rater and the common item effects into account. Although valuable, these models quickly become numerically demanding if the number of items and number of subjects increases. Therefore, in this paper, we propose a new approach that takes the different dependencies in the data into account in a similar way as the existing models, but which can be estimated in a numerically less demanding way. The outline of this paper is as follows: Below we review the current modeling approach and discuss the numerical challenges that motivated the development of our model. Next, as our modeling approach builds upon these models, we start by formally presenting the hierarchical rater model by Patz et al. (2002), the hierarchical rater latent class model by DeCarlo et al. (2011), and the generalized rater model by Wang et al., (2014). Next, we derive our new model, the hierarchical rater thresholds model, and we present an application to real data from the Trends in International Mathematics and Science Study (TIMMSS) comprising ratings of 10 items by 4 raters for 150 subjects. We end with a general discussion.

2 Current Models

As mentioned above, effort has been devoted to develop suitable models for local dependencies due to multiple ratings of the same responses. Among these approaches are the IRT model for multiple raters (Verhelst & Verstralen, 2001), the rater bundle model (Wilson & Hoskens, 2001), the hierarchical rater model (Casabianca, Junker, & Patz, 2016; Patz, Junker, Johnson, & Mariano, 2002), the hierarchical rater latent class model (DeCarlo, Kim, & Johnson, 2011), and the generalized rater model (Wang, Su, & Qiu, 2014). The models are similar in that they account for the common rater effect by rater parameters similarly as in the facet model. The models differ however in the way that they account for the common item effect, see Table 1. The rater bundle model is the most different from the other approaches. In this model, the common item effects are introduced by means of a fixed linear interaction term between pairs of raters of the same item. The interaction effect accounts for the degree to which two raters agree more than what would have been expected on the basis of the student proficiency and the raters' severities.

Table 1: Current modeling approaches to account for the common item effect in rating data.

Model	Source	Common item effect	Estimation
Facet model	Linacre (1989)	-	CML, MML
Rater bundle model	Wilson & Hoskens (2001)	fixed	MML
IRT model for multiple raters	Verhelst & Verstralen (2001)	random, continuous	-
Hierarchical rater model	Patz, et al. (2002)	random, categorical	MCMC
Hierarchical latent class model	DeCarlo, Kim, & Johnson (2011)	random, categorical	MML, PME
Generalized rater model	Wang, Su, & Qiu (2014)	random, continuous	MCMC

Note. CML: Conditional Maximum Likelihood; MML: Marginal Maximum Likelihood; MCMC: Markov Chain Monte Carlo; PME: Posterior Mode Estimation.

In the other models, the common item effect is modeled by introducing a separate latent variable for each item, denoted the ideal ratings. The most important difference between the models in Table 1 is that the hierarchical rater model and the hierarchical rater latent class model treat these ideal ratings as categorical variables (see also DeCarlo, 2005) while in the IRT model for multiple raters and the generalized raters model, the ideal ratings are continuous latent variables. Besides these differences, the models differ in the number of rater parameters. That is, all models, except the facet model and the IRT model for multiple raters, include an additional rater parameter besides the rater severity. The so-called rater variability parameter models the variability in the ratings by the raters. That is, some raters can be more variable in their ratings as compared to other raters. In these models, the rater severity and the rater variability can be either rater specific, or item and rater specific. In the hierarchical rater latent class model by DeCarlo et al. (2011) the rater parameters are specified in a somewhat different way with rater severities being item, rater, and response category depended, and the rater variability being rater and item depended. Note that DeCarlo et al. (2011) use respectively 'response criteria' and 'detection parameters' to refer

to these parameters respectively. We will explicate this difference later in the formal model presentation.

2.1 Estimation Challenges

The models in Table 1 differ in the estimation procedures used to apply the models. That is, while the hierarchical rater model and the generalized rater model rely on Markov Chain Monte Carlo (MCMC) estimation, the hierarchical rater latent class model, the rater bundle model, and the facet model rely mainly on Marginal Maximum Likelihood (MML). Although certainly feasible, both MCMC and MML have their computational challenges with respect to models with a large number of latent variables. As all models but the rater bundle and the facet model include a separate latent variable for each item, the number of latent variables increases rapidly for an increasing number of items. In such high dimensional models, MCMC will become slow and MML may even become unfeasible (see e.g., Wood et al., 2002; although more efficient MML algorithms have been proposed by e.g., Cai, 2010b). In the case of the rater bundle model, as noted by Patz et al. and Wilson and Hoskens (2001), MML estimation becomes computationally infeasible for an increasing number of items and/or raters due to the rapid increase in the number of interaction terms that are needed. In addition, for the IRT model for multiple raters no estimation algorithm has been developed yet as the maximization of the marginal likelihood function is not easy as discussed in Verhelst and Verstralen (2001).

2.2 The Hierarchical Rater Thresholds Model

As discussed above, both MML and MCMC have their challenges when it comes to datasets with an increasing number of items. Therefore, if the number of items is large, a practical model is desirable that both accounts for the variability due to multiple raters and multiple items, but which can also be estimated in large datasets in a numerically efficient way. Therefore, in this paper, we present such an approach. We refer to this approach as 'The Hierarchical Rater Thresholds Model' as the model is a hierarchical model and accounts for differences across raters by explicitly separating the raters' effects from the item effects on the threshold parameters of the categorical observed variables.

The hierarchical rater threshold model draws from the hierarchical rater model by Patz et al (2002), the hierarchical rater latent class model by DeCarlo et al., (2011), and the generalized rater model by Wang et al. (2014). In addition, it utilizes Thurstone's model for categorical judgement (1928; see Torgerson, 1958) in the rater part of the model, and it includes the IRT model for multiple raters by Verhelst and Verstralen (2002) as a special case. The most important features of the hierarchical rater thresholds model are: 1) The model includes less latent variables as compared to the generalized rater model while still accounting for both variability due to difference in items and differences in raters; 2) The model is formulated in such a way that the parameters can be estimated using Weighted Least Squares (WLS; Muthén, 1984), an estimation procedure that is more robust to an increasing numbers of latent variables in terms of convergence and computation time as compared to MCMC and MML; and 3) Similarly to the generalized rater model by Wang et al. (2014), the proposed model is a generalized linear latent variable model (Bartholomew, Knott, & Moustaki, 2011; Moustaki & Knott, 2000; Skrondal, & Rabe-Hesketh, 2004) which provides possibilities to fit the model in standard and flexible software packages like Mplus (Muthén & Muthén, 2007), Lisrel (Jöreskog, & Sörbom , 2001), Amos (Arbuckle, 1997), Mx (Neale, Boker, Xie, & Maes, 2006), SAS (SAS Institute, 2011), and OpenMX (Boker et al., 2010).

3 Formal Models for Raters and Items

3.1 The Hierarchical Rater Model

Level 1. On the first level of the hierarchical rater model the observed ordinal rating, X_{pir}, of student $p = 1, \ldots, N$ on item $i = 1,\ldots,n$ by rater $r = 1,\ldots,R$, are linked to the ordinal ideal ratings, ξ_{pi}, by a normal model, that is, the probability that X_{pir} equals c is given by

$$P\left(X_{pir} = c \middle| \xi_{pi}\right) \propto \exp\left\{-\frac{[c-(\xi_{pi}+\varphi_{ir})]^2}{2\psi_{ir}^2}\right\} \text{ for } c = 0, \ldots, C-1 \quad (1)$$

where $\propto$ denotes 'proportionally to' and φ_{ir} and ψ_{ir} are respectively the item specific rater severity and rater scale parameters (or rater variability parameter ψ^2_r).[1] A lenient

1 The original parameterization uses ψ_r to represent the rater specific variance, however, to enable a comparison to the hierarchical rater threshold model later, we use ψ_r as scale parameter with variability ψ^2_r.

rater who scores on average more towards the upper end of the scale – as compared to the other raters – will be characterized by a large positive φ_{ir}. A severe rater who scores on average more in the lower end of the scale – as compared to the other raters – will be characterized by a large negative φ_{ir}. In addition, a rater who shows more variability in the scores as compared to the other raters will be characterized by a larger ψ_{ir}. Note that these rater effects may be item dependent, that is, a rater can be more lenient or variable on one item and less lenient or variable on another item.

Level 2. At the next level, the ordinal ideal ratings, ξ_{pi}, are linked to the continuous latent student proficiency variable, q_p, using a (generalized) partial credit model, that is, the probability that ξ_{pi} equals k is given by

$$P(\xi_{pi} = k|\theta_p) = \frac{\exp[\sum_{s=o}^{k}(\alpha_i\theta_p - \beta_i) - \gamma_{is}]}{\sum_{t=0}^{K-1}\exp[\sum_{s=0}^{t}(\alpha_i\theta_p - \beta_i) - \gamma_{is}]} \text{ for } k = 0, \ldots, K-1 \quad (2)$$

where α_i is an item discrimination parameter, b_i is the general difficulty of item i (some items are harder than other items) and γ_{jk} is the difficulty of category k with respect to this general difficulty level. At the top level of the model, a normal distribution function for q_p is specified.

The hierarchical rater model above is attractive as it accounts for both the variability in the data due to common raters and variability due to common items. In addition, parameters φ_{ir} and ψ_{ir} enable to quantify rater characteristics which might be helpful to assess rater reliability and to judge individual ratings by a given rater.

3.2 The Hierarchical Rater Latent Class Model

Level 1. The hierarchical rater latent class model is a latent class version of the model above. That is, at level 1, the observed data X_{pir} are connected to the categorical ideal ratings ξ_{pi} by using a graded response model for categorical latent variables:

$$\text{logit}[P(X_{pir} \geq c|\xi_{pi})] = \lambda_{ir}\xi_{pi} - \tau_{irc}$$

Note that although DeCarlo et al. (2011) explicitly leave open the link function, here we focus on the logit link. In the model, λ_{ir} is a rater discrimination parameter (referred to as 'detection parameter' by DeCarlo et al., 2011) which can be item specific, and τ_{irc} are item, rater, and response category specific threshold parameters (referred to as 'relative criteria' by DeCarlo et al., 2011). Note that the model is general, containing many parameters, but both λ_{ir} and τ_{irc} can be constraint to only contain rater, item, or category effects.

Level 2. At level 2 of the hierarchical rater latent class model, a generalized partial credit model is specified similar as in Eq. 2 above, but with category specific difficulty parameters, β_{ik}, instead of a separate overall difficulty β_i and a category parameter γ_{ik}, that is:

$$P(\xi_{pi} = k|\theta_p) = \frac{\exp[\sum_{s=0}^{k}\alpha_i\theta_p - \beta_{is}]}{\sum_{t=0}^{K-1}\exp[\sum_{s=0}^{t}\alpha_i\theta_p - \beta_{is}]} \text{ for } k = 0, \ldots, K-1 \quad (3)$$

3.3 The Generalized Rater Model

Both hierarchical rater models above are based on two categorizations, one on the rater level and one on the observed variable level. This complicates model estimation as each category on each level of the model is associated with separate parameters. An important feature of the generalized rater model by Wang et al. (2014) is that it assumes the ideal scores, ξ_{pi}, to be continuous variables with a normal distribution. Due to this assumption, the model becomes numerically less complex as there are less parameters involved at the latent level (i.e., in the model by Wang, the distribution of latent variable ξ_{pi} can be characterized by a mean and variance parameter, while in the hierarchical rater model separate parameters for all K - 1 latent categories need to be estimated). Besides these continuous latent ideal score variables, the model includes a separate normally distributed latent variable ω_{pr} for each rater with its the mean equal to the item-invariant rater severity, $\mu_\omega = \varphi_r$, and the variance equal to the item-invariant rater specific variability, $\sigma^2_\omega = \psi^2_r$. The resulting model is then given by

$$\text{logit}[P(X_{pir} \geq c|\xi_{pi}, \omega_r, \theta_p)] = \alpha_i\theta_p - \beta_i - \gamma_{ic} - \xi_{pi} - \omega_{pr} \quad (4)$$

where all other parameters have similar interpretations as in the above. Note that the model does not include the hierarchical structure that is characteristic of the hierarchical rater model because there are no higher-order random person effects that are nested in lower-order random person effects. However, the model does account for the local dependencies in the data with ξ_{pi} accounting for the dependencies due to the raters rating the same items, and ω_{pr} accounting for the dependencies due to the items being rated by the same raters.

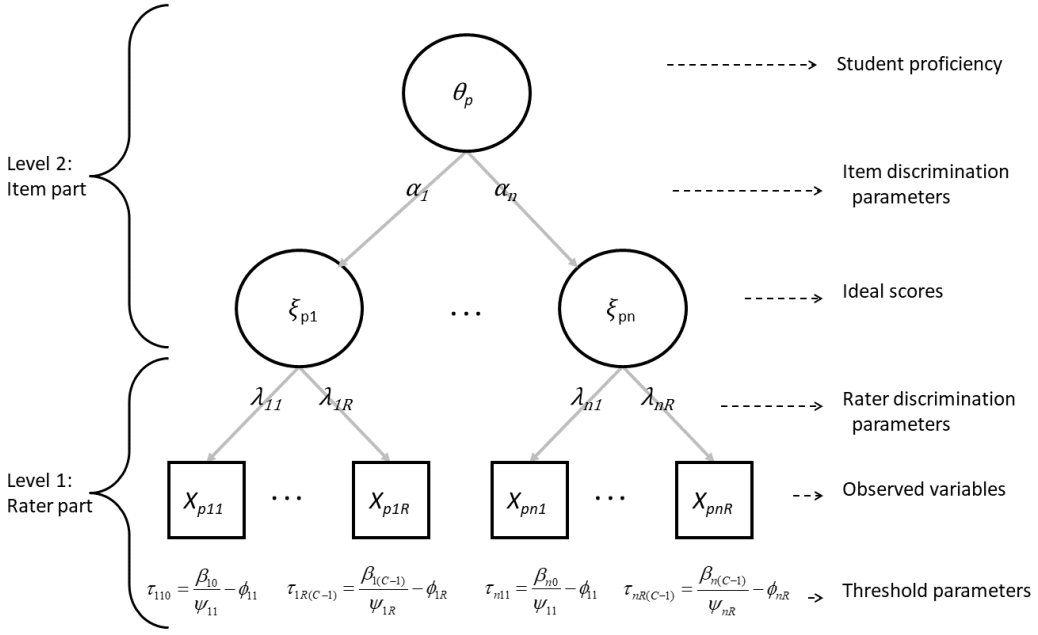

Figure 1: A graphical representation of the hierarchical rater thresholds model.

4 The Hierarchical Rater Thresholds Model

The model we propose in this paper, the hierarchical rater thresholds model, has a hierarchical structure similar to the hierarchical rater model by Patz et al. (2002) and DeCarlo et al. (2011). In addition, the rater thresholds model assumes continuous ideal scores ξ_{pi} similarly as in the generalized rater model by Wang et al. (2014). However, here, we will treat the observed ratings as if they arose from the categorization of the ideal ratings at increasing thresholds. This idea is motivated by Thurstone's model for categorical judgement (1928; see Torgerson, 1958). The resulting models are all special cases in the item factor analysis framework (Wirth & Edwards, 2007) and can therefore be applied using WLS estimation. See Figure 1 for a graphical representation of the hierarchical rater threshold model. Below we describe the model.

4.1 Level 1: Thurstone's model for categorical judgement

Thurstone's model for categorical judgement postulates that categorical judgements in X_{pir} arise by categorization of an underlying normally distributed continuum at increasing thresholds. Here, we use this idea by assuming that the ideal scores ξ_{pi} represent the continuous score that item i by student p should have obtained in the case that an unambiguous and objective continuous scoring rule was available. These continuous scores ξ_{pi} are then internally categorized by the raters at rater and item specific thresholds. Thus, the model postulates that observed ratings of the same item by the same student may differ depending on the rater as raters differ in the thresholds they use to categorize the continuous ideal scores ξ_{pi}, that is

$$X_{pir} = c \quad \text{if } \tau_{irc} < \xi_{pi} < \tau_{ir(c+1)} \text{ for } c = 0, ..., C - 1 \qquad (5)$$

where τ_{irc} denotes the c-th threshold of rater r on item i with $\tau_{i0r} = -\infty$ and $\tau_{iCr} = \infty$. Note that the thresholds do not depend on the student p as we assume, as explicitly discussed in Verhest and Verstralen (2001), that the raters solely base their assessment of item i by student p on the quality of the response and not on other indications for that specific student proficiency (as might for instance happen when the raters are also the students' teachers such that they have an a priori expectation about the students proficiency).

The thresholds in Eq. 5 do not yet include separate item and rater parameters. That is, irrespective of the rater, on some items it may be harder to obtain a high rating because the item is more difficult. In addition, irrespective of the item, it may be harder to obtain a higher rating for some raters because these raters are less

lenient. Therefore, we separate item and rater effects on the thresholds by using

$$\tau_{irc} = \frac{\beta_{ic}}{\psi_{ir}} - \varphi_{ir} \tag{6}$$

where β_{ic} is the difficulty for the c-th category of item i, and φ_{ir} and ψ_{ir} are the item specific rater severity and rater scale parameter (or rater variability ψ^2_r) similarly to the hierarchical rater model in Eq. 1. For reasons of identification, φ_{ir} and ψ_{ir} should be fixed to 0 and 1 for all items for an arbitrarily chosen rater to provide a reference point. Note that φ_{ir} and ψ_{ir} are identified as they account for item and rater specific departures from the main item effect in β_{ic}. Parameters φ_{ir} and ψ_{ir} can be tested to be item-invariant, that is, $\varphi_{ir} = \varphi_r$ and $\psi_{ir} = \psi_r$. However, these items cannot be specified as rater-invariant as $\tau_{irc} = \frac{\beta_{ic}}{\psi_i} - \varphi_i$ will result in an unidentified model as parameters φ_i and ψ_i can be absorbed in β_{ic}. Therefore, in the model specification in Eq. 6, φ_{ir} and ψ_{ir} can truly be seen as rater parameters and not as item parameters. In addition, β_{ic} can truly be seen as item parameters and as not rater parameters.

The way we apply Thurstone's model here is different from the so-called underlying variable approach that is applied in item factor analysis and item response theory (Wirth & Edwards, 2007; Takane & De Leeuw, 1987; Samejima, 1969). Specifically, in the underlying variable approach, a separate variable is assumed to underlie each X_{pir} while, here, we assume a common underlying variable for all r (i.e., ξ_{pi}). Differences in X_{pir} across raters arise due to the differences in the thresholds and not due to differences in the underlying variable.

As the ideal scores are assumed to be normal, applying the categorization scheme in Eq. 5 results in the following model for the probability of a score in category c or higher

$$\text{probit}\left[P\left(X_{pir} \geq c \middle| \xi_{pi}\right)\right] = \lambda_{ir}\xi_{pi} - \tau_{irc} \tag{7}$$

with the thresholds τ_{irc} given by Eq. 6 and with λ_{ir} being an item specific rater discrimination parameter. Note that in this model, the thresholds (and thus φ_{ir} and ψ_{ir}) are on a probit scale. The item specific discrimination parameter can be identified by fixing $\lambda_{ir} = 1$ for all items (i) of an arbitrary rater (τ), or by fixing $\sigma^2_{\xi i} = 1$ for all items (i). Parameter λ_{ir} can be constraint to only reflect a rater effect ($\lambda_{ir} = \lambda_r$), an item effect ($\lambda_{ir} = \lambda_i$), or a common effect ($\lambda_{ir} = \lambda$) which is equivalent to omitting λ_{ir} from Eq. 7 (as the common effect λ can be absorbed in $\sigma^2_{\xi i}$). All these extension can be combined with item specific or item invariant rater parameters for τ_{irc} in Eq. 6. We illustrate this in the real data application. A final note is that the model above is an equivalent to a graded response model (Samejima, 1969) with linear constraints on the item category parameter. In addition, for $C = 2$, $\varphi_{ir} = \varphi_r$, and $\psi_{ir} = 1$, the model simplifies to the model by Verhelst and Verstralen (2001).

4.2 Level 2: A linear latent variable model

Now that the ideal ratings, ξ_{pi} in Eq. 5 are defined and linked to the observed data, they can be submitted to an appropriate measurement model. Here we use the linear common factor model as proposed by Mellenbergh (1994) for continuous item scores, that is,

$$\xi_{pi} = \alpha_i\theta_p + \varepsilon_{pi} \tag{8}$$

where α_i is the item discrimination parameter and ε_{pi} is the first-order item specific residual with variance $\sigma^2_{\varepsilon i}$. Note that the regression does not include an intercept as this parameter is not identified in single group applications. To finalize the model specification, at the top of the model we assume – similarly as in the hierarchical model – that the student proficiency parameter is (standard) normally distributed. For reasons of identification, $VAR(\theta_p)$ should be fixed to 1 or α_i should be fixed to 1 for an arbitrary item.

4.3 Estimation of the Hierarchical Rater Thresholds Model

The full hierarchical rater thresholds model is given by Eq. 6, Eq. 7, and Eq. 8 with a normal distribution for q_p. It is a hierarchical model with first-order factors, ξ_{pi}, second-order factor q_p and constraints on the item category parameters. As all relations in this full model are generalized linear, the model is a member of the generalized linear latent variable modeling framework (Bartholomew, Knott, & Moustaki, 2011; Moustaki & Knott, 2000; Skrondal, & Rabe-Hesketh, 2004). An appealing property of the hierarchical rater thresholds model being a generalized linear latent variable model is that we can estimate the parameters by WLS estimation (see Muthén, 1984). As discussed above, WLS is a suitable and numerically less demanding estimation procedure in the case of an increasing number of latent variables. As in typical models for raters, the number of random effects increases rapidly if the number of items increases. In WLS, the following fitting function is minimized over $\boldsymbol{\sigma}$, the vector of model implied polychoric means and correlations

(which are a function of the free model parameters in Eq. 6, Eq. 7, and Eq. 8):

$$F_{WLS} = (\boldsymbol{s} - \boldsymbol{\sigma})^T \boldsymbol{W}^{-1} (\boldsymbol{s} - \boldsymbol{\sigma})$$

where $\boldsymbol{s}$ is a vector of the estimated polychoric means and correlations of the observed data and $\boldsymbol{W}$ is the covariance matrix of these estimates, $\mathbf{V}$. Matrix $\mathbf{W}$ can become large for an increasing number of observed variables. As $\mathbf{W}$ has to be inverted during estimation, the true benefits of WLS are obtained if $\mathbf{W}$ is taken to be the diagonal of $\mathbf{V}$ (see e.g., Li, 2016; Flora, & Curran, 2004). In that case, the standard errors of the model parameters and the χ^2 goodness-of-fit statistic need to be corrected for the loss of information due to the neglect of the off-diagonal elements of $\mathbf{V}$ during estimation (see Muthén, du Toit, & Spisic, 1997).

As in WLS, expectations are taken over the latent variables ξ_{pi} and θ_p in vector in $\boldsymbol{\sigma}$, these latent variables do not have to be estimated (as in MCMC) or marginalized out (as in MML). The estimation procedure is therefore numerically less demanding, even if there are many latent ξ_{pi} variables (see Muthén, Muthén & Asparouhov, 2005). In a simulation study, Beauducel and Herzberg (2006) showed that a sample size of 250 subjects is enough to fit models up to 8 latent variables and 40 ordinal variables with 2, 3, 4, 5, or 6 point scales. To compare: Wood et al (2002) discuss that more than 5 latent variables is already infeasible for MML (i.e., MML using conventional Gauss-Hermite quadrature approximation; the maximum number of latent variables can be increased by using more efficient approximations as proposed by e.g., Cai, 2010b, but still the estimation can quickly become –at least practically – infeasible).

5 Application

5.1 Data

In this section, we apply the hierarchical rater thresholds model from Eq. 6, Eq. 7, and Eq. 8 to a real data set pertaining mathematics ability. Specifically, the data were collected within the Trends in International Mathematics and Science Study (TIMMSS) in 2007 in Turkey. The data comprise ratings from four raters of 10 items by 150 8th grade students. The raters are all middle school teachers. The items were rated on a 6-point scale. The data were collected in a fully crossed design with all students responding to all items which in turn were rated by all teachers.

Although the data come from a large-scale international assessment, the size of the dataset is admittedly not extremely large (10 items, 4 raters, and 150 subjects). However, we note that the dataset is large enough to demonstrate the value of our approach. That is, in the present dataset MML based approaches are already (practically) infeasible as the model contains 10 latent variables for the common item effects. In addition, as described below, estimation time for the hierarchical rater model by MCMC is already five times longer than the estimation time of our WLS based approach.

5.2 Models

To the data, we fit various versions of the hierarchical rater threshold model with different structures for τ_{irc} and λ_{ir} in Eq. 6. In addition, for a comparison, we fit two rater thresholds models that are based on the facet model in which we omit the common item effects, ξ_{pi}, from the model similarly as in the original facet model by Linacre (1989). That is, we consider a model without ξ_{pi} but with item specific rater parameters, φ_{ir} and ψ_{ir}, and a model without ξ_{pi} but with item invariant rater thresholds, φ_r and ψ_r. Finally, we also consider (different versions of) the hierarchical rater latent class model by DeCarlo (2005) and the hierarchical rater model by Patz et al. (2002).

5.3 Estimation

The hierarchical rater threshold models (including the facet model version) are fit in Mplus (Muthén & Muthén, 2007) using WLS estimation with a diagonal weight matrix $\mathbf{W}$ (also referred to as "diagonally weighted least squares") and a correction on the standard errors and the χ^2 goodness of fit statistic (Muthén et al., 1997). The script to fit the full model in the case of 4 raters and 10 items is available in the Supplementary Materials. As changing this script to a desired number of items or raters can be cumbersome, we wrote an R-script that can be used to generate Mplus-scripts for any number of items and any number of raters, and for all special cases of the model as discussed in this paper. The R-script is available on MASKED.

To assess model fit we use the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993), the Comparative Fit Index (CFI; Bentler & Bonett, 1980) and the Tucker-Lewis Index (TLI; Bentler & Bonett, 1980). According to the guidelines by Schermelleh-Engel, Moosbrugger, and Müller (2003), the RMSEA indicates an acceptable fit for values between 0.08 and 0.05, and a

good fit for values smaller than 0.05. For the CFI and TLI values between 0.95 and 0. 97 indicate acceptable model fit, and values above 0.97 indicate good model fit.

The hierarchical rater model by Patz et al. (2002) is estimated using MCMC estimation in the Immer R-package (Robitzsch & Steinfeld, 2018a; see also Robitzsch & Steinfeld, 2018b; we used 5000 iterations with 1000 burnin). In addition, the hierarchical rater model by DeCarlo (2005) is estimated using maximum likelihood in the R-package Sirt (Robitzsch, 2020; see also Robitzsch & Steinfeld, 2018b). Model fit for the latter model is assessed using the AIC, BIC, CAIC, and AICc fit indices.

6 Results

6.1 Estimation time

To illustrate the efficiency of the present model, we first compared estimation time between the full hierarchical rater thresholds model estimated using WLS and the hierarchical rater model estimated using MCMC estimation as discussed above. The rater threshold model took 55 seconds to estimate on an average laptop, while the hierarchical model took 4 minutes and 38 seconds.[2] Of course, the estimation time of the hierarchical rater model depends on the number of iterations, but we think it does illustrate the difference between both estimation approaches.

6.2 Modeling results: Rater threshold models

The model fit results concerning the rater threshold models are in presented in Table 2. As can be seen, all hierarchical rater thresholds models have an acceptable fit according to the RMSEA with values between 0.058 and 0.79. According to the CFI and TLI, all rater thresholds models fit well with values close to 0.98/0.99. On the contrary, the facet version of the rater threshold model fits poorly. The facet model with item specific rater parameters has a CFI value of 0.911, a TLI value of 0.916, and a RMSEA value of 0.178. The item invariant model also fits poorly with a CFI value of 0.908, a TLI value of 0.919, and a RMSEA value of 0.175.

Parameter estimates for the rater parameters, φ_{ir} and ψ_{ir} in the different rater threshold models are in Table 4, 5, and 6 for Models 1a to 2d from Table 2. Note that Models 3a to 3d do not have separate rater parameters (i.e., $\varphi_{ir} = \varphi = 0$ and $\psi_{ir} = \psi = 1$ for identification reasons in these models). As can be seen from Table 4 and 5, estimates for φ_{ir} and ψ_{ir} hardly differ across Models 1a to 1d. That is, the exact configuration of λ_{ir} does not importantly affect the rater severity and variability estimates, at least for this dataset. However, it can be seen that within a rater, φ_{ir} and ψ_{ir} differ across items. For instance, in Table 4 it can be seen that the estimated severity of rater 2 equals -0.445 (SE: 0.234) on item 5 and 0.535 (SE: 0.079) on item 8. Similar differences are notable for ψ_{ri} in Table 5

Thus, the rater threshold model clearly outperforms the facet model, which was to be expected as the facet model does not take the common item effects into account. However, for the different rater threshold models, it is more difficult to select to best fitting model as all fit indices (especially the TLI and CFI) are very close. Therefore, the question arises which model to choose. As can be seen in Table 2, have restrictions on φ_{ir} and ψ_{ir} generally decreases model fit. That is, all models with restrictions on φ_{ir} and ψ_{ir} fit less well as compared to their corresponding model without restrictions on φ_{ir} and ψ_{ir}. With respect to the restrictions on λ_{ir} is can be seen that have item specific (λ_i) or rater specific (λ_r) parameters generally decreases model fit as compared to an unrestricted λ_{ir}. However, having a common parameter, $\lambda_{ir} = \lambda$ is associated with the best fit. Thus statistically, one would choose the model with an unrestricted φ_{ir} and ψ_{ir} and a common parameter for λ as this model is –at least statistically – the best fitting model. However, as the model fit is close, in practice one would want to take practical and substantive considerations into account as well. In addition, one should keep in mind that the fit measures that we used (CFI, TLI, and RMSEA) tell something about the model implied polychoric correlations. For instance, a difference in RMSEA of 0.002 between Model A and B indicates that Model B is better in describing the polychoric correlation matrix than Model A. Thus in practice, if models have similar RMSEA values, the question is if these differences (i.e., differences in how well the polychoric correlation matrix is described) are important for the research goal at hand.

6.3 Modeling results: Hierarchical latent class models

See Table 3 for the model fit of different hierarchical latent class models according to the AIC, BIC, CAIC, and AICc. As can be seen, the fit indices are slightly mixed with respect to the model that is indicated as the best

2 Configuration of the 'average laptop' is: Intel Core i5 CPU (2.30 Ghz) with 8Gb RAM memory

Table 2: Model fit results for the hierarchical rater threshold models in the application.

Model	τ_{irc}	λ_{ir}	#par	χ^2	df	CFI	TLI	RMSEA
1a	$\tau_{irc} = \frac{\beta_{ic} - \varphi_{ir}}{\psi_{ir}}$	$\lambda_{ir} = \lambda_{ir}$	160	1262.5	820	0.990	0.990	0.060
1b		$\lambda_{ir} = \lambda_i$	130	1402.0	850	0.987	0.989	0.066
1c		$\lambda_{ir} = \lambda_r$	133	1357.9	847	0.988	0.989	0.063
1d		$\lambda_{ir} = \lambda$	121	1285.5	859	0.990	0.991	0.058
2a	$\tau_{irc} = \frac{\beta_{ic} - \varphi_r}{\psi_r}$	$\lambda_{ir} = \lambda_{ir}$	106	1405.0	874	0.988	0.989	0.064
2b		$\lambda_{ir} = \lambda_i$	76	1541.7	904	0.986	0.988	0.069
2c		$\lambda_{ir} = \lambda_r$	79	1498.4	901	0.986	0.988	0.066
2d		$\lambda_{ir} = \lambda$	67	1405.6	913	0.989	0.990	0.060
3a	$\tau_{irc} = \beta_{ic}$	$\lambda_{ir} = \lambda_{ir}$	100	1622.1	880	0.983	0.985	0.075
3b		$\lambda_{ir} = \lambda_i$	70	1751.9	910	0.981	0.984	0.079
3c		$\lambda_{ir} = \lambda_r$	73	1711.0	907	0.982	0.984	0.077
3d		$\lambda_{ir} = \lambda$	61	1559.3	919	0.985	0.988	0.068

Note. $\tau_{irc} = \frac{\beta_{ic} - \varphi_i}{\psi_i}$ is not considered as it is not identified (see main text). In addition, '#par' denotes: number of parameters in the model.

Table 3: Model fit results for the hierarchical rater latent class models in the application.

τ_{irc}	λ_{irc}	#par	AIC	BIC	CAIC	AICc
$\tau_{irc} = \tau_{ic}$	$\lambda_{ir} = \lambda_{ir}$	150	**10833**	11285	11435	*
	$\lambda_{ir} = \lambda_i$	120	12033	12394	12514	13034
	$\lambda_{ir} = \lambda_r$	114	10951	11294	11408	11700
	$\lambda_{ir} = \lambda$	111	12036	12370	12481	12690
$\tau_{irc} = \tau_{rc}$	$\lambda_{ir} = \lambda_{ir}$	120	10705	**11066**	**11186**	11706
	$\lambda_{ir} = \lambda_i$	90	10891	11162	11252	11169
	$\lambda_{ir} = \lambda_r$	84	10924	11177	11261	11144
	$\lambda_{ir} = \lambda$	81	10942	11186	11267	**11138**
$\tau_{irc} = \tau_c$	$\lambda_{ir} = \lambda_{ir}$	105	10914	11230	11335	11420
	$\lambda_{ir} = \lambda_i$	75	12085	12311	12386	12239
	$\lambda_{ir} = \lambda_r$	69	11123	11331	11400	11244
	$\lambda_{ir} = \lambda$	66	12167	12365	12431	12273

Note. Best values of the fit indices are in bold face.
*: as the number of parameters equals the sample size, the AICc cannot be calculated for this model

fitting model. That is, the AIC prefers the model with $\tau_{irc} = \tau_{ic}$ and $\lambda_{ir} = \lambda_{ir}$ while the AICc prefers the model with $\tau_{irc} = \tau_{rc}$ and $\lambda_{ir} = \lambda$. The BIC and CAIC both agree in that the model with $\tau_{irc} = \tau_{rc}$ and $\lambda_{ir} = \lambda_{ir}$ is the best fitting model. Therefore, we accept this model as the final hierarchical latent class model.

6.4 Modeling results: Comparison

A direct comparison between the rater threshold model and the hierarchical rater latent class model is not possible as estimation of the former is not based on a likelihood function. As a result, for the rater threshold model, no

Table 4: Parameter estimates (standard errors) for the rater severity, φ_{ri} across Model 1a to Model 1d and the hierarchical rater model (HRM).

		Rater Threshold Model				HRM
r	*i*	*Model 1a*	*Model 1b*	*Model 1c*	*Model 1d*	
1	1	0*	0*	0*	0*	-0.708 (0.074)
1	2	0*	0*	0*	0*	-0.720 (0.115)
1	3	0*	0*	0*	0*	-0.635 (0.129)
1	4	0*	0*	0*	0*	-0.972 (0.159)
1	5	0*	0*	0*	0*	0.467 (0.126)
1	6	0*	0*	0*	0*	-0.753 (0.102)
1	7	0*	0*	0*	0*	-1.159 (0.122)
1	8	0*	0*	0*	0*	-0.932 (0.131)
1	9	0*	0*	0*	0*	-0.368 (0.125)
1	10	0*	0*	0*	0*	-0.285 (0.139)
2	1	0.544 (0.056)	0.544 (0.056)	0.544 (0.056)	0.544 (0.056)	0.425 (0.094)
2	2	0.160 (0.074)	0.160 (0.074)	0.160 (0.074)	0.160 (0.074)	-0.181 (0.141)
2	3	0.524 (0.079)	0.524 (0.079)	0.524 (0.079)	0.524 (0.079)	0.571 (0.13)
2	4	0.400 (0.062)	0.400 (0.062)	0.400 (0.062)	0.400 (0.062)	0.307 (0.174)
2	5	-0.445 (0.234)	-0.445 (0.234)	-0.445 (0.234)	-0.445 (0.234)	0.119 (0.094)
2	6	0.586 (0.079)	0.586 (0.079)	0.586 (0.079)	0.586 (0.079)	0.438 (0.134)
2	7	0.494 (0.058)	0.494 (0.058)	0.494 (0.058)	0.494 (0.058)	0.159 (0.134)
2	8	0.535 (0.079)	0.535 (0.079)	0.535 (0.079)	0.535 (0.079)	0.238 (0.137)
2	9	0.247 (0.058)	0.247 (0.058)	0.247 (0.058)	0.247 (0.058)	0.188 (0.099)
2	10	0.447 (0.062)	0.447 (0.062)	0.447 (0.062)	0.447 (0.062)	0.731 (0.135)
3	1	0.526 (0.062)	0.526 (0.062)	0.526 (0.062)	0.526 (0.062)	0.755 (0.171)
3	2	0.263 (0.078)	0.263 (0.078)	0.263 (0.078)	0.263 (0.078)	0.117 (0.150)
3	3	0.408 (0.070)	0.408 (0.070)	0.408 (0.070)	0.408 (0.070)	0.397 (0.152)
3	4	0.426 (0.065)	0.426 (0.065)	0.426 (0.065)	0.426 (0.065)	0.531 (0.192)
3	5	0.220 (0.140)	0.220 (0.140)	0.220 (0.140)	0.220 (0.140)	0.445 (0.133)
3	6	0.735 (0.085)	0.735 (0.085)	0.735 (0.085)	0.735 (0.085)	1.002 (0.148)
3	7	0.557 (0.067)	0.557 (0.067)	0.557 (0.067)	0.557 (0.067)	0.357 (0.099)
3	8	0.641 (0.088)	0.641 (0.088)	0.641 (0.088)	0.641 (0.088)	0.611 (0.178)
3	9	0.295 (0.055)	0.295 (0.055)	0.295 (0.055)	0.295 (0.055)	0.403 (0.106)
3	10	0.625 (0.080)	0.625 (0.080)	0.625 (0.080)	0.625 (0.080)	1.483 (0.178)
4	1	0.435 (0.049)	0.435 (0.049)	0.435 (0.049)	0.435 (0.049)	0.112 (0.081)
4	2	0.203 (0.062)	0.203 (0.062)	0.203 (0.062)	0.203 (0.062)	0.158 (0.112)
4	3	0.491 (0.085)	0.491 (0.085)	0.491 (0.085)	0.491 (0.085)	0.158 (0.083)
4	4	0.419 (0.060)	0.419 (0.060)	0.419 (0.060)	0.419 (0.060)	0.371 (0.188)
4	5	0.235 (0.146)	0.235 (0.146)	0.235 (0.146)	0.235 (0.146)	0.23 (0.086)
4	6	0.632 (0.082)	0.632 (0.082)	0.632 (0.082)	0.632 (0.082)	0.629 (0.147)
4	7	0.317 (0.056)	0.317 (0.056)	0.317 (0.056)	0.317 (0.056)	-0.167 (0.146)

Continued **Table 4:** Parameter estimates (standard errors) for the rater severity, φ_{ri} across Model 1a to Model 1d and the hierarchical rater model (HRM).

		Rater Threshold Model				HRM
r	*i*	*Model 1a*	*Model 1b*	*Model 1c*	*Model 1d*	
4	8	0.516 (0.081)	0.516 (0.081)	0.516 (0.081)	0.516 (0.081)	0.244 (0.127)
4	9	0.267 (0.057)	0.267 (0.057)	0.267 (0.057)	0.267 (0.057)	0.201 (0.09)
4	10	0.347 (0.068)	0.347 (0.068)	0.347 (0.068)	0.347 (0.068)	0.254 (0.105)

*: This parameter is fixed for identification purposes.

Table 5: Parameter estimates (standard errors) for the rater scale, ψ_{ri} across Model 1a to Model 1d and the hierarchical rater model (HRM).

		Rater Threshold Model				HRM
r	*i*	*Model 1a*	*Model 1b*	*Model 1c*	*Model 1d*	
1	1	1*	1*	1*	1*	0.747 (0.061)
1	2	1*	1*	1*	1*	0.802 (0.081)
1	3	1*	1*	1*	1*	1.139 (0.092)
1	4	1*	1*	1*	1*	1.376 (0.120)
1	5	1*	1*	1*	1*	0.841 (0.065)
1	6	1*	1*	1*	1*	0.945 (0.081)
1	7	1*	1*	1*	1*	1.057 (0.104)
1	8	1*	1*	1*	1*	1.145 (0.093)
1	9	1*	1*	1*	1*	1.090 (0.087)
1	10	1*	1*	1*	1*	1.196 (0.095)
2	1	1.061 (0.106)	1.061 (0.106)	1.061 (0.106)	1.061 (0.106)	0.668 (0.057)
2	2	1.047 (0.157)	1.047 (0.157)	1.047 (0.157)	1.047 (0.157)	1.043 (0.093)
2	3	2.991 (0.568)	2.991 (0.568)	2.991 (0.568)	2.991 (0.568)	0.895 (0.065)
2	4	1.279 (0.192)	1.279 (0.192)	1.279 (0.192)	1.279 (0.192)	1.012 (0.099)
2	5	0.659 (0.115)	0.659 (0.115)	0.659 (0.115)	0.659 (0.115)	0.633 (0.051)
2	6	2.620 (0.481)	2.620 (0.481)	2.621 (0.481)	2.620 (0.481)	0.813 (0.087)
2	7	1.586 (0.184)	1.586 (0.184)	1.586 (0.184)	1.586 (0.184)	0.696 (0.110)
2	8	2.890 (0.514)	2.890 (0.514)	2.890 (0.514)	2.890 (0.514)	0.711 (0.089)
2	9	1.480 (0.198)	1.480 (0.198)	1.480 (0.198)	1.480 (0.198)	0.786 (0.091)
2	10	1.926 (0.253)	1.926 (0.253)	1.926 (0.253)	1.926 (0.253)	0.950 (0.080)
3	1	1.172 (0.105)	1.172 (0.105)	1.172 (0.105)	1.172 (0.105)	1.268 (0.100)
3	2	1.042 (0.161)	1.042 (0.161)	1.042 (0.161)	1.042 (0.161)	1.295 (0.098)
3	3	2.197 (0.344)	2.197 (0.344)	2.197 (0.344)	2.197 (0.344)	1.132 (0.080)
3	4	1.614 (0.260)	1.614 (0.260)	1.614 (0.260)	1.614 (0.260)	1.192 (0.134)
3	5	1.357 (0.297)	1.357 (0.297)	1.357 (0.297)	1.357 (0.297)	0.777 (0.075)
3	6	2.739 (0.541)	2.739 (0.541)	2.739 (0.541)	2.739 (0.541)	0.853 (0.068)

Continued **Table 5:** Parameter estimates (standard errors) for the rater scale, ψ_{ri} across Model 1a to Model 1d and the hierarchical rater model (HRM).

		Rater Threshold Model				HRM
r	*i*	*Model 1a*	*Model 1b*	*Model 1c*	*Model 1d*	
3	7	2.122 (0.316)	2.122 (0.316)	2.122 (0.316)	2.122 (0.316)	0.543 (0.109)
3	8	3.954 (0.870)	3.954 (0.870)	3.954 (0.870)	3.954 (0.870)	0.759 (0.111)
3	9	1.310 (0.171)	1.310 (0.171)	1.310 (0.171)	1.310 (0.171)	0.862 (0.069)
3	10	2.902 (0.472)	2.902 (0.472)	2.902 (0.472)	2.902 (0.472)	1.264 (0.098)
4	1	0.973 (0.086)	0.973 (0.086)	0.973 (0.086)	0.973 (0.086)	0.502 (0.046)
4	2	1.207 (0.159)	1.207 (0.159)	1.207 (0.159)	1.207 (0.159)	0.846 (0.140)
4	3	4.019 (0.803)	4.019 (0.803)	4.019 (0.803)	4.019 (0.803)	0.562 (0.051)
4	4	1.585 (0.244)	1.585 (0.244)	1.585 (0.244)	1.585 (0.244)	0.938 (0.150)
4	5	1.357 (0.291)	1.357 (0.291)	1.357 (0.290)	1.357 (0.291)	0.498 (0.052)
4	6	2.705 (0.491)	2.705 (0.491)	2.705 (0.491)	2.705 (0.491)	0.856 (0.084)
4	7	1.686 (0.172)	1.686 (0.172)	1.686 (0.172)	1.686 (0.172)	1.118 (0.104)
4	8	3.289 (0.608)	3.289 (0.608)	3.289 (0.608)	3.289 (0.608)	0.831 (0.118)
4	9	1.677 (0.240)	1.677 (0.240)	1.677 (0.240)	1.677 (0.240)	0.700 (0.105)
4	10	2.340 (0.310)	2.340 (0.310)	2.340 (0.310)	2.340 (0.310)	0.389 (0.086)

*: This parameter is fixed for identification purposes.

likelihood based model fit statistics can be calculated. However, an indirect comparison is possible as the best fitting models can be compared in terms of the effects that are included. That is, the final rater threshold model was argued to be the model with item and rater specific severity and scale parameters φ_{ir} and ψ_{ir} and with rater and item invariant parameter λ. For the hierarchical rater latent class model, the final model includes $\tau_{irc} = \tau_{rc}$ and $\lambda_{ir} = \lambda_{ir}$. Both models thus agree on rater and item specific effects, however, in the latent class model, the rater effects are modeled in the thresholds, while the items effects are modeled in the rater discrimination parameter λ_{ir}. This is in accordance with how DeCarlo et al. (2011) proposed the compare the hierarchical rater latent class model to the hierarchical rater model by Patz et al. (2002). That is, De Carlo et al compare τ_{irc} and λ_{ir} from the hierarchical latent class model to respectively φ_r and ψ_r from the hierarchical rater model.

A comparison in terms of parameter estimates between the rater threshold model and the hierarchical rater latent class model is challenging as the threshold parameters τ_{irc} are restricted in a different way (but see DeCarlo et al., 2011 for a possible graphical approach). Here, we focus on a comparison between the rater threshold model and the hierarchical rater model by Patz et al. (2002) as both models include comparable rater severity and rater scale parameters φ_{ir} and ψ_{ir}. See below.

6.5 Modeling results: Hierarchical rater models

Parameter estimates for the rater parameters, φ_{ir} and ψ_{ir} from the hierarchical rater model are in Table 4 and 5. Table 6 contains the rater parameters in a model with item invariant rater parameter φ_r and ψ_r. Note that in the original hierarchical rater model by Patz et al (2002), ψ^2_{ir} is estimated, but here we focus on ψ_{ir} to facilitate comparison to the rater threshold models.

6.6 Modeling results: Comparison

In comparing the rater parameters φ_{ir} and ψ_{ir} between the hierarchical rater model and the rater threshold model a transformation of the hierarchical rater model parameters is necessary. That is, in the rater threshold model, the rater parameters reflect the differences among the raters relative to rater 1, as for this rater, $\varphi_{i1} = 0$ and $\psi_{i1} = 0$ for identification reasons. Therefore, the comparison between

Table 6: Parameter estimates (standard errors) for the rater severity, φ_r, and the rater scale, ψ_r for Model 2a to Model 2d and the hierarchical rater model (HRM).

		Rater Threshold Model				HRM
Parameter	*r*	*Model 2a*	*Model 2b*	*Model 2c*	*Model 2d*	
φ_r	1	0*	0*	0*	0*	-0.594 (0.035)
	2	0.485 (0.048)	0.485 (0.048)	0.485 (0.048)	0.485 (0.048)	0.364 (0.037)
	3	0.627 (0.061)	0.627 (0.061)	0.627 (0.061)	0.627 (0.061)	0.69 (0.046)
	4	0.498 (0.051)	0.498 (0.051)	0.498 (0.051)	0.498 (0.051)	0.456 (0.034)
ψ_r	1	1*	1*	1*	1*	1.022 (0.025)
	2	1.317 (0.063)	1.317 (0.063)	1.317 (0.063)	1.317 (0.063)	0.824 (0.023)
	3	1.443 (0.066)	1.443 (0.066)	1.443 (0.066)	1.443 (0.066)	1.043 (0.025)
	4	1.418 (0.061)	1.418 (0.061)	1.418 (0.061)	1.418 (0.061)	0.641 (0.025)

*: This parameter is fixed for identification purposes.

the rater parameters in Table 4, 5, and 6 should be based on the ordering of the raters. That is, in Table 5 it can be seen that for rater severity, φ_r, the hierarchical rater model and the rater threshold model result in the same ordering of the raters (i.e., from least to most severe: 1,2,4,3). For the rater scale parameter however, the models differ in their implied ordering. That is, for the rater threshold model the ordering from smallest scale to widest scale is: 1, 2, 4, and 3, while the ordering for the hierarchical rater model is: 4, 2, 1, and 3.

To look more into the relation between the rater parameters from the two models, we transform the hierarchical rater model parameters to reflect rater severity and scale relative to the first rater, that is:

$$\varphi'_{ir} = \varphi_{ir} - \varphi_{i1}$$

and

$$\psi'_{ir} = \psi_{ir}/\psi_{i1}.$$

Next, for raters 2, 3, and 4, we correlate these transformed parameters from the hierarchical rater model to the parameter from the rater threshold model. In addition, we consider the observed marginal means ($M_{ir} = MEAN_p(X_{pir})$) and standard deviations ($S_{ir} = SD_p(X_{pir})$) of the rater's item ratings. Similar to above, we also transformed these statistics into a measure relative to rater 1 (M^*_{ir} and S^*_{ir}). Table 7 depicts the correlations between the relative hierarchical rater parameters, the relative rater statistics, and the rater parameters from the rater threshold model. As can be seen, the correlations between M^*_{ir} and φ^*_{HRM} are generally large for all three raters (around 0.970). The

Table 7: For rater 2 to 4: the correlations between the relative rater mean (M^*_r) and the relative rater standard deviation (SD^*_r) of the observed item ratings (X_{pir}), and the relative rater parameters from the hierarchical rater model (φ^*_{HRM} and ψ^*_{HRM}) and the rater parameters from the rater threshold model (φ_{RTM} and ψ_{HRM}) over items .

r		M^*_r	SD^*_r	φ^*_{HRM}	ψ^*_{HRM}	φ_{RTM}	ψ_{HRM}
2	M^*_r	1					
	SD^*_r	0.129	1				
	φ^*_{HRM}	0.972	0.257	1			
	ψ^*_{HRM}	-0.287	-0.129	-0.216	1		
	φ_{RTM}	0.936	0.269	0.963	-0.143	1	
	ψ_{HRM}	0.497	0.609	0.585	-0.326	0.650	1
3	M^*_r	1					
	SD^*_r	0.146	1				
	φ^*_{HRM}	0.961	0.100	1			
	ψ^*_{HRM}	-0.193	-0.363	-0.098	1		
	φ_{RTM}	0.853	0.001	0.891	-0.24	1	
	ψ_{HRM}	0.546	0.358	0.569	-0.521	0.760	1
4	M^*_r	1					
	SD^*_r	-0.121	1				
	φ^*_{HRM}	0.964	-0.017	1			
	ψ^*_{HRM}	0.287	-0.159	0.438	1		
	φ_{RTM}	0.729	-0.171	0.648	-0.075	1	
	ψ_{HRM}	0.306	0.469	0.273	-0.311	0.640	1

Note. See the main text for an explanation about the relative statistics and parameters.

Table 8: For each rater, the correlations between the rater mean (M_r) and the rater standard deviation (SD_r) of the observed item ratings (X_{pir}), and the rater parameters φ_{ir} and ψ_{ir} from the hierarchical rater model (HRM).

r		M_r	SD_r	φ_{HRM}	ψ_{HRM}
1	M_r	1			
	SD_r	-0.477	1		
	φ_{HRM}	0.716	-0.180	1	
	ψ_{HRM}	-0.119	0.134	-0.305	1
2	M_r	1			
	SD_r	-0.634	1		
	φ_{HRM}	0.634	-0.419	1	
	ψ_{HRM}	-0.652	0.583	0.036	1
3	M_r	1			
	SD_r	-0.726	1		
	φ_{HRM}	0.604	-0.589	1	
	ψ_{HRM}	-0.420	0.040	0.218	1
4	M_r	1			
	SD_r	-0.593	1		
	φ_{HRM}	0.415	-0.066	1	
	ψ_{HRM}	-0.375	0.390	-0.121	1

correlation between $M^{'}_{ir}$ and φ_{RTM} is about as large for rater 2 (0.936), but somewhat smaller for rater 2 (0.853) and rater 3 (0.729). The correlation between $\varphi^{'}_{HRM}$ and φ_{RTM} is large for raters 2 and 3 (around 0.900), but somewhat smaller for rater 4 (0.648). Next, the correlation between SD_{ir} and $\psi^{'}_{HRM}$ is small and negative (between -0.129 and -0.363) while the correlation between SD_{ir} and ψ_{RTM} is larger and positive (between 0.358 and 0.609). The correlation between $\psi^{'}_{HRM}$ and ψ_{RTM} is also negative (between -0.311 and -0.521). Thus overall, it seems that both $\varphi^{'}_{HRM}$ and φ_{RTM} capture marginal relative mean differences between raters well, with the hierarchical rater model being somewhat better, while the rater threshold model captures relative differences in the marginal standard deviation better. As we focused on the transformed parameters of the hierarchical rater model relative to rater 1, we also provide the correlations between the untransformed hierarchical rater parameters and M_{ir} and S_{ir}, see Table 8. As can be seen, the S_{ir} now correlates positively with the rater scale parameter φ_{ir} although this correlation is small for rater 1 and 3.

In conclusion, φ_{ir} from the hierarchical rater model and from the rater threshold model can both be used to quantify relative differences between raters in the marginal means. The scale parameter ψ_{ir} from the hierarchical rater model is not suitable to quantify relative differences between raters in the marginal standard deviations while ψ_{ir} from the rater threshold model performs better in this respect. Thus, the rater threshold should ideally be preferred if the aim is to quantify relative differences between raters. However, if the aim is to quantify differences between items within a rater, the hierarchical rater model is better suitable as for the rater threshold model, only relative differences between raters are quantified.

7 Discussion

In this paper we presented a method to infer student proficiency from items rated by multiple raters. As we relied on WLS estimation, we argued that our approach is suitable for large scale educational settings where multiple latent variables are needed to account for violations of local independence. Our argument is mainly a pragmatic one. There may be good reasons to choose the hierarchical rater model or a Bayesian approach over the present approach. For instance, contrary to the Bayesian hierarchical rater model by Patz et al. (2002), our approach is sensitive to small sample sizes. That is, for decreasing sample sizes, parameter estimates in the present model will become biased or even infeasible. In such cases, a Bayesian approach like the hierarchical rater model is desired. In addition, there may be substantive/conceptual reasons to choose for the hierarchical rater model due to the categorical nature of the ideal scores in this model. For instance, as argued by Mariano & Junker (2007), if the common item effects are treated as continuous variables, these cannot be interpreted as "ideal scores" because the scale differs from those of the observed ratings.

Despite sample size considerations, there are two more trade-offs associated with our approach (see also Muthén, Muthén, & Asparouhov, 2015). First, our approach is more sensitive to the number of variables in the analysis as compared to MCMC or MML. That is, as the estimation of our model is based on the asymptotic covariance matrix of the polychoric means and correlations, it becomes more and more challenging to estimate the full matrix if there are both many raters and many items. However, as in practice the number of raters is commonly limited, the present approach will be suitable in many cases. Second, contrary to MML and MCMC approaches, our approach is a limited information approach which only uses the information from the first two polychoric moments in the data. However, the effects of neglecting higher-

order moments is generally considered to be small (see Christoffersson, 1975).

The main difference between the present approach and other approaches (e.g., Patz et al., 2002; Wilson & Hoskens, 2001) is that the rater effects are treated as effects on the thresholds at which an underlying continuum is categorized by the raters. We chose this approach both for pragmatic reasons (i.e., to be in the generalized linear latent variable modeling framework) and substantive reasons (i.e., to facilitate the psychological interpretation of the parameters using Thurstone's law of categorical judgment, 1928). This choice implies that the rater effect is treated as a fixed effect. This different from, for instance, Snijders and Bosker (1999, Chapter 11) and Wang et al. (2014). The question whether the rater effect is a random or a fixed effect is probably an empirical one. If the raters are truly selected randomly from a larger pool of raters, it might be best to account for this source of variation. However, we think that in practice, raters are selected according to specific criteria (e.g., availability, rating skill, acquaintance with the topic to be rated, etc.) such that the assumption of fixed rater effects is defendable.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47–76.

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

Arbuckle, J. L. (1997). Amos (version 3.61) [Computer software]. Chicago, IL: Small Waters.

SAS Institute Inc. (2011). *SAS/STAT software: Release 9.3*. Cary, NC: SAS Institute, Inc.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011*). Latent Variable Models and Factor Analysis: A Unified Approach*. UK: John Wiley & Sons, Ltd.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In E M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (chap. 17–20), Reading, MA: Addison Wesley.

Boker, S., Neale. M.C., Maes, H.H., Wilde, M., Spiegel, M., Brick, T., et al. (2010) OpenMx: an open source extended structural equation modeling framework. *Psychometrika 76,* 306–317.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153–168.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136 – 162). Thousand Oaks, CA7 Sage Publications.

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins Monro algorithm. *Psychometrika*, *75*, 33–57.

Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. In *Handbook of Item Response Theory, Volume One* (pp. 477–494). Chapman and Hall/CRC.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5–32.

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

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*, 333–356.

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

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56–70.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466.

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

Jöreskog, K. G. & Sörbom, D. (2001). *LISREL user's guide*. Chicago: Scientific Software International.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*, 936–949.

Linacre, J. M. (1989). *Many-faceted Rasch Measurement.* Chicago: MESA Press.

Lord, F. M. (1952). *A Theory of Test Scores*. New York: Psychometric Society.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS – a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337.

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine, 28*, 3049–3067.

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

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response assessments*. Unpublished doctoral dissertation. Pittsburgh: Carnegie Mellon University.

Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test

items. *Journal of Educational and Behavioral Statistics, 32*, 287–314.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223–236.

Mislevy, R.J., & Bock, R.D. (1989). A hierarchical item-response model for educational testing. In R.D. Bock (Eds.), *Multilevel Analysis of Educational Data* (pp. 57–74). San Diego, CA: Academic Press.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika, 65*, 391–411.

Muthén BO, du Toit SHC, & Spisic D. (1997). *Robust Inference using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes*. Unpublished manuscript. Retrieved from https://www.statmodel.com/download/Article_075.pdf

Muthén, L.K., & Muthén, B.O. (2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2015). Estimator choices with categorical outcomes. Unpublished manuscript, retrieved from: https://www.statmodel.com/download/EstimatorChoices.pdf

Neale, M.C., Boker, S.M., Xie, G., & Maes, H.H. (2006). *Mx: Statistical Modeling, 7th ed*. VCU, Department of Psychiatry, Richmond.

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*, 341–384.

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

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York: Springer. Sas Institute. (2011). *SAS/STAT 9.3 user's guide*. SAS Institute.

Robitzsch, A. (2020). sirt: Supplementary Item Response Theory Models. R package version 3.9-4. https://CRAN.R-project.org/package=sirt

Robitzsch, A., & Steinfeld, J. (2018a). immer: Item response models for multiple ratings. R package version 1.1-35. https://CRAN.R-project.org/package=immer

Robitzsch, A., & Steinfeld, J. (2018b). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling,* 60, 101–139.

Samejima, F. (1969). *Estimation of Ability using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: The Psychometric Society.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23–74.

Sireci, S. G., Wainer, H., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall/CRC.

Snijders, T., & Bosker, R. (2012). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. London etc.: Sage Publishers.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Thurstone, L. L. (1928). The measurement of opinion. *Journal of Abnormal and Social Psychology, 22*, 415–430.

Torgerson, W.S. (1958). Theory and methods of scaling. Oxford, England: Wiley.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus Frequentist inference. In H. Hoijtink, I. Klugkist, and P. A. Boelen (Eds.), *Bayesian Evaluation of Informative Hypotheses*, pp. 181–207. Springer: New York.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8*, 157–187.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22–29.

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

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*, 283–306.

Wilson, M, & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19,* 51–72.

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.

Wood, R., Wilson, D. T., Gibbons, R. D., Schilling, S. G., Muraki, E., & Bock, R. D. (2002). TESTFACT: Test scoring, item statistics, and item factor analysis. Chicago: Scientific Software International.