



A multidimensional generalized many-facet Rasch model for rubric-based performance assessment

Masaki Uto¹

Received: 1 June 2021 / Accepted: 15 July 2021 / Published online: 30 July 2021
© The Author(s) 2021

Abstract

Performance assessment, in which human raters assess examinee performance in a practical task, often involves the use of a scoring rubric consisting of multiple evaluation items to increase the objectivity of evaluation. However, even when using a rubric, assigned scores are known to depend on characteristics of the rubric's evaluation items and the raters, thus decreasing ability measurement accuracy. To resolve this problem, item response theory (IRT) models that can estimate examinee ability while considering the effects of these characteristics have been proposed. These IRT models assume unidimensionality, meaning that a rubric measures one latent ability. In practice, however, this assumption might not be satisfied because a rubric's evaluation items are often designed to measure multiple sub-abilities that constitute a targeted ability. To address this issue, this study proposes a multidimensional IRT model for rubric-based performance assessment. Specifically, the proposed model is formulated as a multidimensional extension of a generalized many-facet Rasch model. Moreover, a No-U-Turn variant of the Hamiltonian Markov chain Monte Carlo algorithm is adopted as a parameter estimation method for the proposed model. The proposed model is useful not only for improving the ability measurement accuracy, but also for detailed analysis of rubric quality and rubric construct validity. The study demonstrates the effectiveness of the proposed model through simulation experiments and application to real data.

Keywords Multidimensional item response theory · Many-facet Rasch model · Rater bias · Bayesian modeling · Educational/psychological measurement

Communicated by Kazuo Shigemasu.

✉ Masaki Uto
uto@ai.lab.uec.ac.jp

¹ The University of Electro-Communications, Tokyo, Japan

1 Introduction

In various assessment fields, performance assessment, in which raters assess examinee outcomes or processes for a performance task, has attracted much attention as a way to measure practical and higher-order abilities, such as problem-solving, critical reasoning, and creative thinking skills (Mislevy 2018; Zlatkin-Troitschanskaia et al. 2019; Murtonen and Balloo 2019; Palm 2008; Shavelson et al. 2019; Linlin 2019; Hussein et al. 2019; Uto and Okano 2020). Performance assessment has been conducted in various formats, including essay writing, oral presentations, interview examinations, and group discussions.

Performance assessment often involves the use of a scoring rubric that consists of multiple evaluation items to increase the objectivity of evaluation. However, even when using a rubric, assigned scores are known to depend on the characteristics of the rubric's evaluation items and the raters, which decreases the ability measurement accuracy (Deng et al. 2018; Hua and Wind 2019; Myford and Wolfe 2003; Nguyen et al. 2015; Rahman et al. 2020; Uto et al. 2020; Uto and Ueno 2018). Therefore, to improve measurement accuracy, ability estimation that considers the effects of these characteristics is needed.

For this reason, item response theory (IRT) models that can estimate examinee abilities while considering the effects of these characteristics have been proposed (Uto and Ueno 2018; Linacre 1989; Jin and Wang 2018; Wilson and Hoskens 2001; Shin et al. 2019). One representative model is the many-facet Rasch model (MFRM) (Linacre 1989), which has been applied to various performance assessments (Linlin 2019; Hua and Wind 2019; Deng et al. 2018; Chan et al. 2017; Tavakol and Pinner 2019; Kaliski et al. 2013). However, the MFRM makes strong assumptions, for example all raters having the same consistency level and all evaluation items having the same discriminatory power, even though these assumptions rarely hold in practice (Myford and Wolfe 2003; Jin and Wang 2018; Patz et al. 2002; Uto and Ueno 2020; Soo Park and Xing 2019). To relax these assumptions, several extensions of the MFRM have been recently proposed (Uto and Ueno 2018; Jin and Wang 2018; Shin et al. 2019; Uto and Ueno 2020). These IRT models are known to measure abilities with higher accuracy compared with simple scoring methods based on point totals or averages (Uto and Ueno 2016, 2020).

The IRT models assume unidimensionality, meaning that a rubric measures one latent ability. However, this assumption might not hold in practical rubric-based performance assessment because evaluation items in a rubric are often designed to measure multiple sub-abilities that comprise a targeted ability. Applying unidimensional IRT models to data with multidimensional ability scales deteriorates model fitting and ability measurement accuracy (Hutten 1980).

Multidimensional IRT models are used to measure examinee abilities on a multidimensional ability scale and are well known in objective testing scenarios (Reckase 2009). Traditional multidimensional IRT models, however, have no rater parameters, which prevents not only estimation of examinee ability while considering the effects of rater characteristics, but also direct application to rubric-based performance assessment data.

To resolve the abovementioned problems, the present study proposes a new multidimensional IRT model that incorporates parameters that take into account the characteristics of both the rubric’s evaluation items and the raters. Specifically, the proposed model is formulated by extending the generalized MFRM (Uto and Ueno 2020), which is one of the latest MFRM extension models, based on the approach of the multidimensional generalized partial credit model (GPCM) (Yao and Schwarz 2006). The study adopts a Bayesian estimation method for the proposed model called the No-U-Turn (NUT) Hamiltonian Monte Carlo (HMC), which is a state-of-the-art Markov chain Monte Carlo (MCMC) algorithm (Hoffman and Gelman 2014). As outlined in Fig. 1, the proposed model has the following advantages.

1. The proposed model allows for estimation of examinee abilities while considering the effects of the rubric’s evaluation items and the raters simultaneously, which improves model fitting and ability measurement accuracy.
2. Examinee abilities can be assessed on a multidimensional ability scale that is assumed under the rubric’s evaluation items.
3. The model provides characteristic parameters for the rubric’s evaluation items while removing the effects of the raters and the examinees. The parameters enable us to conduct detailed analysis of the rubric’s characteristics and construct validity.

We demonstrate the effectiveness of the proposed model through simulation experiments and application to actual data.

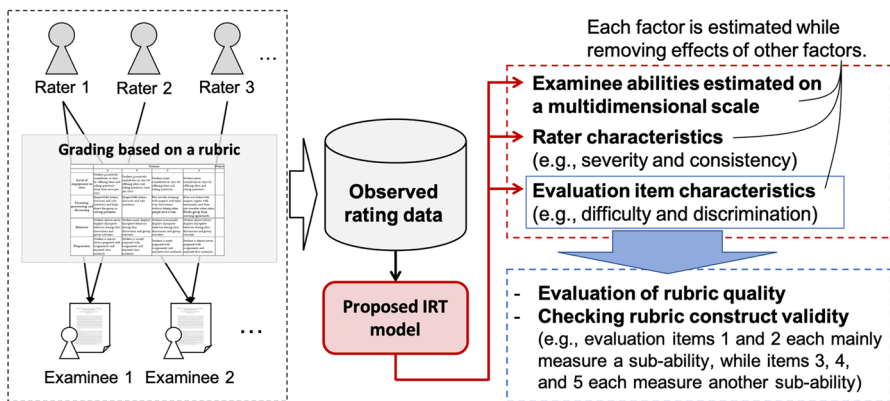


Fig. 1 Outline of an application of the proposed model

2 Rating data from rubric-based performance assessment

This study assumes situations where examinee performance on a given task is assessed by multiple raters using a scoring rubric consisting of multiple evaluation items. Thus, obtained rubric-based performance assessment data X are defined as follows:

$$X = \{x_{ijr} | x_{ijr} \in \mathcal{K} \cup \{-1\}, i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\}, \quad (1)$$

where x_{ijr} is a rating assigned to the performance of examinee $j \in \mathcal{J} = \{1, 2, \dots, J\}$ by rater $r \in \mathcal{R} = \{1, 2, \dots, R\}$ based on evaluation item $i \in \mathcal{I} = \{1, 2, \dots, I\}$ in the rubric. $\mathcal{K} = \{1, 2, \dots, K\}$ represents rating categories, and $x_{ijr} = -1$ represents missing data.

This study aimed to estimate examinee ability from data X by using IRT.

3 Item response theory

With widespread adoption of computer testing, there has been an increase in the use of IRT (Lord 1980), a test theory based on mathematical models. In objective testing contexts, IRT generally defines the relationship between observed examinee responses to test items and latent examinee ability variables using latent variable models, so-called IRT models. IRT models give the probability of an item response as a function of the examinee's latent ability and the item's characteristic parameters. IRT offers the following benefits:

1. Examinee ability can be estimated while considering test item characteristics.
2. Examinee responses to different test items can be assessed on the same scale.
3. Statistically sound methods for handling missing data can be easily applied.

IRT has traditionally been applied to test items for which responses can be scored as correct or incorrect, such as multiple-choice items. In recent years, however, there have been attempts to apply polytomous IRT models to performance assessments (Reise and Revicki 2014). Well-known IRT models that are applicable to ordered-categorical data such as rubric-based performance assessment data include the rating scale model (RSM) (Andrich 1978), partial credit model (PCM) (Masters 1982), and GPCM (Muraki 1997).

The GPCM gives the probability that examinee j receives score k for test item i as:

$$P_{ijk} = \frac{\exp \sum_{m=1}^k [1.7\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [1.7\alpha_i(\theta_j - \beta_i - d_{im})]}, \quad (2)$$

where θ_j is the latent ability of examinee j , α_i is a discrimination parameter for test item i , β_i is a difficulty parameter for item i , and d_{im} is a step parameter denoting difficulty of transition between scores $m - 1$ and m for item i . Here, $d_{i1} = 0$, $\sum_{m=2}^K d_{im} = 0$, and a normal prior for the ability θ_j are assumed for model

identification. The constant 1.7 is often used to make the model similar to the normal ogive function.

The PCM is a special case of the GPCM, where $\alpha_i = 1.0$ for all the items. The RSM is a special case of the PCM, where $d_{im} = d_m$ for all the items. Here, d_m denotes difficulty of transition between categories $m - 1$ and m .

Such traditional IRT models are applicable to two-way data consisting of examinees \times test items. However, they cannot be directly applied to rubric-based performance assessment data comprising examinees \times raters \times evaluation items, even if we regard the test item parameters as the evaluation item parameters. To address this problem, IRT models that can consider these characteristics jointly have been proposed (Uto and Ueno 2018; Linacre 1989; Jin and Wang 2018; Wilson and Hoskens 2001; Shin et al. 2019). Note that some such IRT models originally consider characteristics of performance tasks and raters assuming three-way data consisting of examinees \times raters \times performance tasks. However, this study assumes that IRT models are applied to rubric-based performance assessment data by regarding the performance task parameters as the evaluation item parameters.

4 IRT models for performance assessment

The most widely used IRT model for performance assessment is the MFRM (Linacre 1989). There are several variants of the MFRM (Myford and Wolfe 2004; Eckes 2015), but the most representative modeling defines the probability that $x_{ijr} = k \in \mathcal{K}$ as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k [\theta_j - \beta_i - \beta_r - d_m]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\theta_j - \beta_i - \beta_r - d_m]}, \tag{3}$$

where β_i is a difficulty parameter for evaluation item i , and β_r is the severity of rater r . For model identification, $\sum_{r=1}^R \beta_r = 0$, $d_1 = 0$, $\sum_{m=2}^K d_m = 0$, and a normal prior for the ability θ_j are assumed.

A unique feature of this MFRM is that it is defined by the fewest parameters in existing IRT models for performance assessment. The accuracy of parameter estimation generally increases as the number of parameters per data point decreases (Uto and Ueno 2016, 2020; van der Linden 2016). Thus, this MFRM can estimate model parameters from a small dataset more accurately than can other models, resulting in higher ability measurement accuracy if it fits well to given data (Uto and Ueno 2018; van der Linden 2016).

By contrast, the MFRM relies on the following assumptions.

1. All evaluation items have the same discriminatory power.
2. All raters have the same assessment consistency.
3. All raters share an equal interval rating scale.

However, these assumption are rarely satisfied in practice (Deng et al. 2018; Myford and Wolfe 2003; Jin and Wang 2018; Patz et al. 2002; Uto and Ueno 2020; Soo Park and Xing 2019; Elliott et al. 2009). Violation of these assumptions would decrease model fitting and ability measurement accuracy (Uto and Ueno 2018).

To relax these assumptions, various extensions of the MFRM have recently been proposed (Uto and Ueno 2018; Jin and Wang 2018; Shin et al. 2019; Uto and Ueno 2020; Eckes 2015). This study introduces the generalized MFRM (Uto and Ueno 2020), which relaxes all three assumptions simultaneously. This model provides the probability that $x_{ijr} = k$ as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k [1.7\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [1.7\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]}, \quad (4)$$

where α_i is a discrimination parameter for evaluation item i , α_r is the consistency of rater r , and d_{rm} is a step parameter denoting severity of rater r of transition from rating category $m - 1$ to m . The rater-specific step parameter d_{rm} can represent the rating scale for each rater, meaning that the restriction of an equal-interval scale for raters is relaxed. For model identification, $\prod_{i=1}^I \alpha_i = 1$, $\sum_{i=1}^I \beta_i = 0$, $d_{r1} = 0$, $\sum_{m=2}^K d_{rm} = 0$, and a normal prior for the ability θ_j are assumed.

The generalized MFRM can represent various rater effects and evaluation item characteristics, so it is expected to provide better model fitting and more accurate ability measurement than the MFRM, especially when various rater and evaluation item characteristics are assumed (Uto and Ueno 2020). Thus, this study develops a multidimensional model for rubric-based performance assessment based on the generalized MFRM.

Note that there are various approaches to dealing with rater effects, such as hierarchical rater models (Patz et al. 2002; DeCarlo et al. 2011), extensions based on signal detection models (Soo Park and Xing 2019; DeCarlo 2005), rater bundle models (Wilson and Hoskens 2001), and trifactor models (Shin et al. 2019). However, this study focuses on the MFRM-based approach because it is the most popular and traditional approach.

5 Multidimensional item response theory models

In objective testing contexts, the use of multidimensional IRT models that can measure examinee ability in multidimensional space has been widespread (Reckase 2009). Multidimensional IRT models can generally be classified into compensatory and non-compensatory models. Compensatory models assume that an examinee achieves high performance if any one of multiple sub-abilities is sufficiently high, whereas non-compensatory models assume that performance quality depends concurrently on multiple sub-abilities. This study focuses on non-compensatory models for the following reasons. (1) The non-compensatory assumption would be more suitable for performance assessment settings because performing practical complex tasks will generally require multiple abilities, not a single specific ability. (2) Compensatory models require more

complex model formulations, which increases the number of model parameters and makes interpretation and estimation of parameters difficult.

A representative non-compensatory multidimensional IRT model for polytomous data is the non-compensatory multidimensional GPCM (Yao and Schwarz 2006). When test item parameters are regarded as evaluation item parameters, the model gives the probability that examinee j obtains score k for evaluation item i as

$$P_{ijk} = \frac{\exp \sum_{m=1}^k \left[1.7 \left(\sum_{l=1}^L \alpha_{il} \theta_{jl} - \beta_i - d_{im} \right) \right]}{\sum_{l=1}^K \exp \sum_{m=1}^l \left[1.7 \left(\sum_{l=1}^L \alpha_{il} \theta_{jl} - \beta_i - d_{im} \right) \right]}, \tag{5}$$

where L indicates the number of assumed ability dimensions, θ_{jl} is the ability of examinee j for dimension $l \in \mathcal{L} = \{1, \dots, L\}$, and α_{il} indicates the discriminatory power of evaluation item i for the l -th ability dimension. For model identification, $d_{i1} = 0$, $\sum_{m=2}^K d_{im} = 0$, and a normal prior for the ability of each dimension θ_{jl} are assumed.

However, such conventional multidimensional IRT models have no rater parameters, which prevents not only estimation of examinee ability while considering the effects of rater characteristics, but also direct application to rubric-based writing assessment data. To address this limitation, this study proposes a new multidimensional IRT model that considers the characteristics of both the rubric’s evaluation items and the raters.

6 Proposed model

The proposed model is formulated as a multidimensional extension of the generalized MFRM based on the multidimensional GPCM approach. Specifically, this model gives the probability that $x_{ijr} = k \in \mathcal{K}$ as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k \left[1.7 \alpha_r \left(\sum_{l=1}^L \alpha_{il} \theta_{jl} - \beta_i - \beta_r - d_{im} \right) \right]}{\sum_{l=1}^K \exp \sum_{m=1}^l \left[1.7 \alpha_r \left(\sum_{l=1}^L \alpha_{il} \theta_{jl} - \beta_i - \beta_r - d_{im} \right) \right]}. \tag{6}$$

The proposed model can estimate examinee ability on a multidimensional scale while considering the characteristics of both the evaluation items and the raters compared with conventional models that cannot consider rater characteristics nor estimate ability on a multidimensional scale. Thus, the proposed model is expected to provide better model fitting and more accurate ability measurement compared with the conventional models if ability multidimensionality and rater effects are assumed to exist. Furthermore, application of the proposed model to the rubric-based writing assessment data provides the various characteristics of each evaluation item and rater, which helps in interpreting the quality of the evaluation items and the raters. Also, the evaluation item’s discrimination parameters α_{il} offers information for interpreting what each ability dimension measures, which makes objective analysis of rubric construct validity possible.

Note that the generalized MFRM incorporates a rater-specific step parameter d_{rm} , assuming that the appearance tendency of each rating category m depends on the raters. In rubric-based performance assessment, however, the appearance tendency of each category is expected to depend more strongly on the rubric's evaluation items than the raters because evaluation criteria for each rating category are defined by rubrics. Therefore, the proposed model defines the step parameter as evaluation item's parameter d_{im} instead of d_{rm} .

6.1 Parameter interpretation

This subsection describes how to interpret parameters in the proposed model.

Figure 2 shows item response surfaces (IRSs) based on Eq. (6) for six pairs of raters and evaluation items with different characteristics, with a fixed number of ability dimensions $L = 2$ and rating categories $K = 4$. The parameters used for the IRSs are shown in Table 1. For example, in Fig. 2a, the IRS uses the parameter values for Evaluation item 1 and Rater 1 in Table 1. The x -axis indicates the ability in the first dimension θ_{j1} , the y -axis indicates the ability in the second dimension θ_{j2} , and the z -axis indicates the probability of P_{ijrk} . The IRSs show that the probability of obtaining higher scores increases as the examinees' abilities increase.

Figure 2b–f shows IRSs in which a specific model parameter is changed from that of Fig. 2a. Thus, how each parameter works is explained below by comparing each figure with Fig. 2a.

Figure 2b shows the IRS in which the discrimination parameter for the second dimension, α_{j2} , is smaller than that shown in Fig. 2a. A change in the response probabilities that arises from a change in the second-dimension ability value becomes smaller than that in Fig. 2a. This means that evaluation items with smaller discrimination for the l -th dimension of α_{jl} do not distinguish the corresponding ability dimension, θ_{jl} , well. It also suggests that the ability dimension an evaluation item mainly measures would be interpreted based on the discrimination parameter, as we explain in the next subsection.

Figure 2c shows the IRS with a higher difficulty parameter, β_i . This IRS shows that an increase in difficulty parameter β_i causes the IRS to shift in the direction of the ability value increase, which reflects that difficulty in obtaining higher scores increases as the difficulty parameter for evaluation items increases.

Figure 2d shows the IRS in which the step difficulty parameters, d_{i2} and d_{i3} , are changed. As the difference $d_{i(m+1)} - d_{im}$ increases, the probability of obtaining category m increases over widely varying ability scales. In Fig. 2d, the probability for rating category 2 increases compared with that in Fig. 2a because $d_{i3} - d_{i2}$ is large, whereas the probability for rating category 3 decreases because $d_{i4} - d_{i3}$ is small.

Figure 2e shows the IRS for a rater with a lower consistency parameter, α_r . Compared with Fig. 2a, the differences in the response probabilities among the categories decrease, which reflects inconsistencies in rater scoring among examinees with the same ability level. In contrast, a higher consistency value α_r results in large differences in the response probabilities among the categories. This observation reflects the tendency of consistent raters to consistently assign the

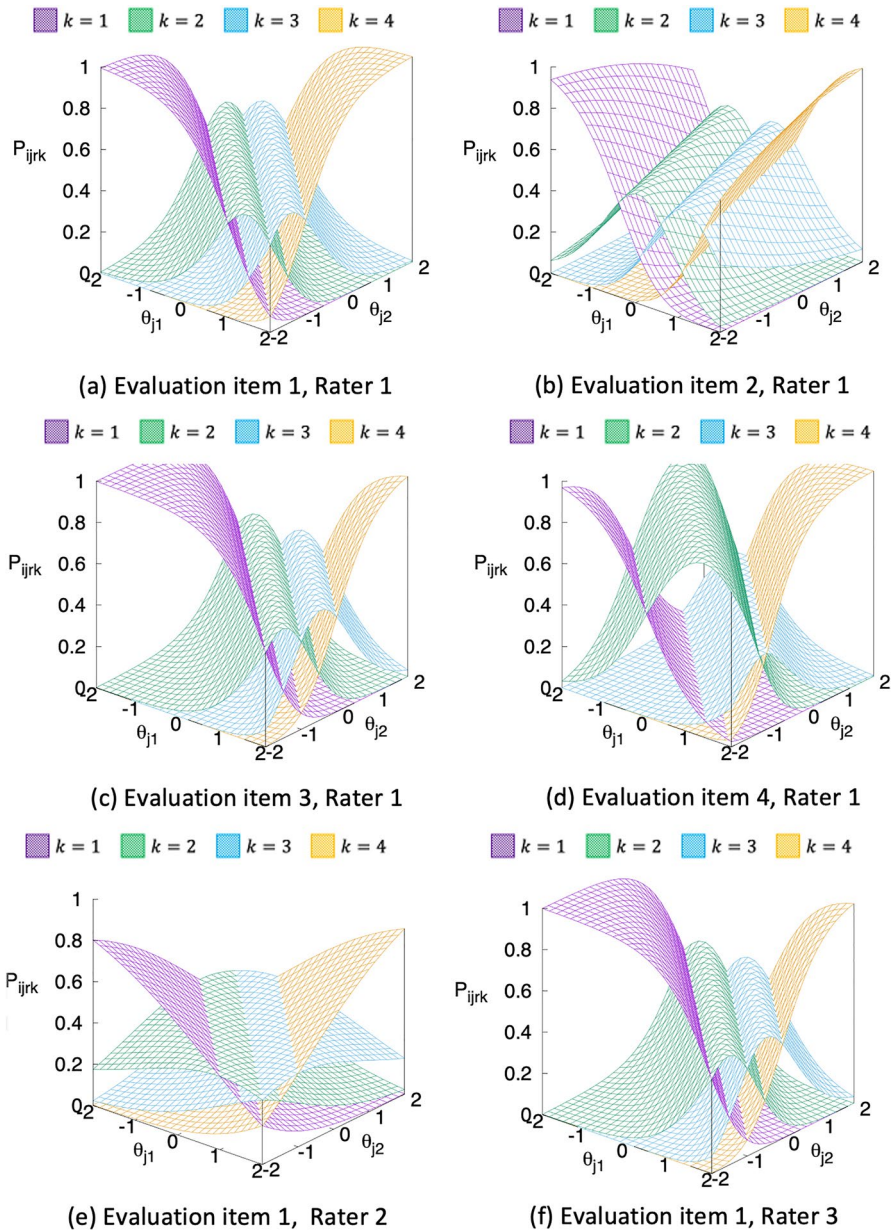


Fig. 2 Item response surfaces for evaluation items and raters with various characteristics

same ratings to examinees with the same ability level and higher ratings to examinees with higher ability levels. This suggests that raters who are more consistent in scoring are generally desirable for accurate ability measurement.

Table 1 Parameters used in Fig. 2

	α_{i1}	α_{i2}	β_i	d_{i2}	d_{i3}	d_{i4}
Evaluation item 1	1.0	1.0	0.0	-1.0	0.0	1.0
Evaluation item 2	1.0	0.3	0.0	-1.0	0.0	1.0
Evaluation item 3	1.0	1.0	1.0	-1.0	0.0	1.0
Evaluation item 4	1.0	1.0	0.0	-2.0	0.8	1.0
	α_r					
	β_r					
Rater 1	1.0		0.0			
Rater 2	0.3		0.0			
Rater 3	1.0		1.0			

Figure 2f shows the IRS for a rater with a high severity parameter, β_r . Compared with Fig. 2a, the IRS shifts in the direction of the increase in ability value as the difficulty parameter for evaluation items increases, reflecting difficulty in assignment of higher ratings by severe raters.

6.2 Interpretation of ability dimensions

As noted above, the discrimination parameter for each dimension of α_{il} offers information for interpreting what each ability dimension measures. Specifically, by analyzing commonality in content among the evaluation items with higher discrimination values for each dimension, we can interpret what each ability dimension mainly measures.

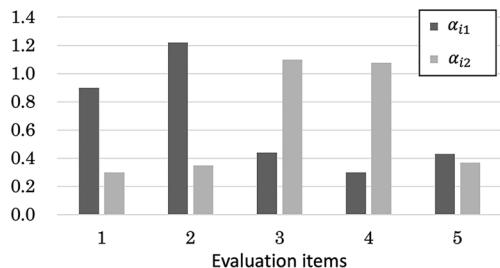
This analysis is explained in greater detail in Fig. 3, which depicts the discrimination parameters for five evaluation items. The horizontal axis indicates the index of the evaluation items, the vertical axis indicates the value of the discrimination parameters, and the colored bars correspond to the discrimination parameter for each dimension. The figure indicates that evaluation items 1 and 2 have larger discrimination values in the first dimension. This suggests that the first dimension mainly relates to a common ability that underlies evaluation items 1 and 2. Thus, by analyzing commonality in content between evaluation items 1 and 2, we can interpret the meaning of the first ability dimension. Similarly, the meaning of the second ability dimension can be interpreted by investigating commonality in content between evaluation items 3 and 4, which have higher discrimination parameters for the second dimension.

6.3 Optimal number of dimensions

In the proposed model, the number of dimensions, L , is a hand-tuned parameter that must be determined in advance. In IRT studies, the optimal number of dimensions is generally explored based on principal component analysis. However, this analysis method is not applicable to the three-way data assumed in the present study.

Dimensionality selection, which is well known in machine learning, can also be considered as a model selection task. The model selection is typically conducted using information criteria, such as the Akaike information criterion (AIC) (Akaike 1974), the Bayesian information criterion (BIC) (Schwarz 1978), the widely

Fig. 3 Discrimination parameters for five evaluation items



applicable information criterion (WAIC) (Watanabe 2010) and the widely applicable Bayesian information criterion (WBIC) (Watanabe 2013). AIC and BIC are applicable when maximum likelihood estimation is used to estimate model parameters, whereas the WAIC and WBIC can be used with Bayesian estimation using MCMC or variational inference methods. With the recent increase in complex statistical and machine learning models, various studies have used WAIC and WBIC because Bayesian estimation tends to provide a more robust estimation for complex models (Vehtari et al. 2017; Almond 2014; Luo and Al-Harbi 2017). Because this study uses a Bayesian estimation based on MCMC, as described in Subsect. 6.5, it uses the WAIC and WBIC to select the optimal number of dimensions for the proposed model. Specifically, the dimensionality that minimizes these criteria is regarded as optimal.

6.4 Model identifiability

The proposed model entails a non-identifiability problem whereby parameter values cannot be uniquely determined because different value sets can give the same response probability. For the proposed model without rater parameters that is consistent with the conventional multidimensional GPCM, parameters are known to be identifiable by assuming a specific distribution (e.g., standard normal distribution) for the ability and constraining $d_{i1} = 0$ and $\sum_{m=2}^K d_{im} = 0$ for each i .

However, in the proposed model, the location for $\beta_i + \beta_r$ is indeterminate, even when these constraints are given, because the response probability with β_i and β_r gives the same value of P_{ijrk} with $\beta'_i = \beta_i + c$ and $\beta'_r = \beta_r - c$ for any constant c (note that $\beta'_i + \beta'_r = (\beta_i + c) + (\beta_r - c) = \beta_i + \beta_r$). Such location indeterminacy can be solved by fixing one parameter or restricting the mean of some parameters (Uto and Ueno 2020; Fox 2010).

There is another indeterminacy of the scale for α_r . Suppose we let the term $\sum_{l=1}^L \alpha_{il} \theta_{jl} - \beta_i - \beta_r - d_{im}$ in Eq. (6) be ξ , then the response probability P_{ijrk} with α_r and ξ gives the same value of P_{ijrk} with $\alpha'_r = \alpha_r c$ and $\xi' = \frac{\xi}{c}$ for any constant c , because $\alpha'_r \xi' = (\alpha_r c) \frac{\xi}{c} = \alpha_r \xi$. Such scale indeterminacy can be removed by fixing one parameter or restricting the product of some parameters (Uto and Ueno 2020; Fox 2010).

This study therefore uses the restrictions $\prod_{r=1}^R \alpha_r = 1$, $\sum_{r=1}^R \beta_r = 0$ for model identification, in addition to $d_{i1} = 0$, $\sum_{m=2}^K d_{im} = 0$, and the standard normal prior for the ability of each dimension θ_{jl} .

6.5 Parameter estimation using MCMC

This subsection describes the parameter estimation method for the proposed model.

Marginal maximum likelihood estimation using an expectation–maximization algorithm is a widely used method for estimating IRT model parameters (Baker and Kim 2004). However, for complex models such as that proposed in this study, expected a posteriori (EAP) estimation, a type of Bayesian estimation, is known to provide more robust estimations (Uto and Ueno 2016; Fox 2010).

EAP estimates are calculated as the expected value of the marginal posterior distribution of each parameter. The posterior distribution in the proposed model is

$$g(\theta_{jl}, \log \alpha_{il}, \log \alpha_r, \beta_i, \beta_r, d_{im} | X) \propto L(X | \theta_{jl}, \log \alpha_{il}, \log \alpha_r, \beta_i, \beta_r, d_{im}) g(\theta_{jl} | \tau_{\theta_{jl}}) g(\log \alpha_{il} | \tau_{\alpha_{il}}) g(\log \alpha_r | \tau_{\alpha_r}) g(\beta_i | \tau_{\beta_i}) g(\beta_r | \tau_{\beta_r}) g(d_{im} | \tau_{d_{im}}), \tag{7}$$

where $\theta_{jl} = \{\theta_{jl} | j \in \mathcal{J}, l \in \mathcal{L}\}$, $\log \alpha_{il} = \{\log \alpha_{il} | i \in \mathcal{I}, l \in \mathcal{L}\}$, $\log \alpha_r = \{\log \alpha_r | r \in \mathcal{R}\}$, $\beta_i = \{\beta_i | i \in \mathcal{I}\}$, $\beta_r = \{\beta_r | r \in \mathcal{R}\}$, and $d_{im} = \{d_{im} | i \in \mathcal{I}, k \in \mathcal{K}\}$. In addition, $g(\mathcal{S} | \tau_s) = \prod_{s \in \mathcal{S}} g(s | \tau_s)$, where \mathcal{S} is a set of parameters, $g(s | \tau_s)$ indicates a prior distribution for parameter s , and τ_s is its hyperparameters. $L(X | \theta_{jl}, \log \alpha_{il}, \log \alpha_r, \beta_i, \beta_r, d_{im})$ is the likelihood that can be calculated as $\prod_{j=1}^J \prod_{i=1}^I \prod_{r=1}^R \prod_{k=1}^K (P_{ijrk})^{z_{ijrk}}$, where z_{ijrk} is a dummy variable that takes 1 if $x_{ijr} = k$, and zero otherwise.

The marginal posterior distribution for each parameter is derived by marginalizing across all parameters except the target parameter. For a complex IRT model, however, it is not generally feasible to derive or calculate the marginal posterior distribution due to high-dimensional multiple integrals. To address this problem, there is widespread use of MCMC, a random sampling-based estimation method, in various fields including IRT studies (Fox 2010; Uto 2019; van Lier et al. 2018; Fontanella et al. 2019; Zhang et al. 2011; Uto et al. 2017; Louvigné et al. 2018; Brooks et al. 2011).

The Metropolis-Hastings-within-Gibbs sampling method (Patz and Junker 1999) is a common MCMC algorithm for IRT models. The algorithm is simple and easy to implement but requires a long time to converge to the target distribution (Hoffman and Gelman 2014; Girolami and Calderhead 2011). As an efficient alternative MCMC algorithm, the NUT sampler (Hoffman and Gelman 2014), a variant of the HMC, has recently been proposed along with a software package called ‘‘Stan’’ (Carpenter et al. 2017), which makes implementation of a NUT-based HMC easy. Thus, there has been recent widespread use of NUT for parameter estimations in various statistical models, including IRT models (Uto and Ueno 2020; Luo and Jiao 2018; Jiang and Carter 2019).

We therefore use a NUT-based MCMC algorithm for parameter estimations in the proposed model. The estimation program was implemented in RStan (Stan Development Team 2018). The developed Stan code is provided in Appendix 1. In this study, the standard normal distribution $N(0.0, 1.0)$ is used as a prior distribution for each parameter: θ_{jl} , $\log \alpha_{il}$, $\log \alpha_r$, β_i , β_r , and d_{im} . Furthermore, the EAP estimates are calculated as the mean of parameter samples obtained from 2000 to 4000 periods.

7 Simulation experiments

7.1 Accuracy of parameter recovery

In this subsection, we describe a parameter recovery experiment for the proposed model through simulations. We conducted the following experiments by

changing the number of examinees, evaluation items, raters, and dimensions to $J \in \{50, 100\}$, $I \in \{5, 15\}$, $R \in \{5, 15\}$, and $L \in \{1, 2, 3\}$, respectively.

1. For J examinees, I evaluation items, R raters, and L dimensions, generate true model parameters randomly from the following distributions, which are the same as the prior distributions used in the MCMC algorithm.

$$\theta_{jl}, \log \alpha_{il}, \log \alpha_r, \beta_i, \beta_r, d_{im} \sim N(0.0, 1.0). \quad (8)$$

The number of rating categories, K , was fixed to 4 to match the condition of the actual data used in a later section.

2. Set skewed discrimination values to the first L evaluation items $i \in \{1, \dots, L\}$ as follows:

$$\begin{cases} \alpha_{il} = 1.5 & i = l \\ \alpha_{il} = 0.2 & i \neq l \end{cases} \quad (9)$$

The necessity of this procedure is explained in the next paragraph.

3. Given the true parameters, generate rating data from the proposed model randomly.
4. Estimate the model parameters from the generated data.
5. Sort the order of the estimated dimensions based on the discrimination parameter estimates. Specifically, using the discrimination parameter estimates for the first L evaluation items, we sorted the dimensions so that $\sum_{i=1}^L \alpha_{il}$ was maximized.
6. Calculate root mean square errors (RMSEs) and biases between the estimated and true parameters.
7. Repeat the above procedure 30 times, then calculate the average values of the RMSEs and biases.

Note that Procedures 2 and 5 are required because the ability dimensions in multidimensional IRT models including the proposed model are exchangeable, meaning that the dimension to which a sub-ability corresponds changes every time the parameter estimation runs. The dimension indeterminacy is caused because interchanging $\alpha_{il}\theta_{jl}$ and $\alpha_{i'l'}\theta_{j'l'}$ ($l \in \mathcal{L}$, $l' \in \mathcal{L}$, $l \neq l'$) results in the same value for the term $\sum_{i=1}^L \alpha_{il}\theta_{jl}$. To appropriately calculate the RMSEs and biases between the estimated and true parameters, this experiment addressed the problem by setting extreme discrimination parameter values for the first L evaluation items in Procedure 2, and by sorting the estimated dimensions based on the discrimination parameter estimates in Procedure 5, as in Martin-Fernandez and Revuelta (2017).

Table 2 shows the RMSE results, which confirm the following tendencies:

1. The RMSEs for ability values tend to decrease as the number of evaluation items and/or raters increases. Similarly, the RMSEs for raters and evaluation item parameters tend to decrease as the number of examinees increases. These ten-

- dencies are caused by the increase in the amount of data per parameter, which is consistent with previous studies (Uto and Ueno 2016, 2018).
2. An increase in the number of dimensions tends to lead to an increase in the RMSEs because the ability and discrimination parameters increase without an increase in the amount of data. This tendency is also consistent with previous research on multidimensional IRT (Martin-Fernandez and Revuelta 2017; Svetina et al. 2017; Kose and Demirtasli 2012).

Moreover, Table 3 shows that the average bias was nearly zero in all cases, indicating no overestimation or underestimation of parameters. We also confirmed that the Gelman–Rubin statistic \hat{R} (Gelman and Rubin 1992; Gelman et al. 2013), a well-known convergence diagnostic index, was less than 1.1 in all cases, indicating that the MCMC runs converged.

From the above, we conclude that the parameter estimation for the proposed model can be appropriately conducted using the MCMC algorithm.

Table 2 RMSE values obtained from the parameter recovery experiment

<i>J</i>	<i>I</i>	<i>R</i>	<i>L</i>	θ_{jl}	α_{il}	α_r	β_i	β_r	d_{im}	
50	5	5	1	0.212	0.160	0.102	0.089	0.079	0.181	
			2	0.337	0.217	0.097	0.139	0.069	0.195	
			3	0.393	0.270	0.131	0.200	0.094	0.213	
		15	1	0.144	0.106	0.131	0.057	0.074	0.111	
			2	0.234	0.145	0.149	0.105	0.076	0.132	
			3	0.285	0.198	0.165	0.146	0.083	0.120	
		15	5	1	0.149	0.152	0.067	0.107	0.039	0.219
				2	0.264	0.154	0.063	0.135	0.042	0.202
				3	0.307	0.167	0.063	0.154	0.047	0.223
	15	5	1	1	0.111	0.101	0.068	0.055	0.041	0.122
				2	0.200	0.121	0.085	0.096	0.043	0.132
				3	0.217	0.131	0.081	0.104	0.046	0.137
			2	1	0.217	0.114	0.090	0.068	0.056	0.135
				2	0.319	0.157	0.083	0.100	0.053	0.146
				3	0.362	0.160	0.092	0.119	0.067	0.158
100	5	5	1	0.131	0.079	0.094	0.043	0.058	0.083	
			2	0.215	0.098	0.089	0.072	0.054	0.093	
			3	0.249	0.116	0.084	0.090	0.056	0.090	
		15	1	0.133	0.094	0.038	0.061	0.031	0.154	
			2	0.238	0.109	0.047	0.089	0.032	0.160	
			3	0.286	0.120	0.043	0.103	0.031	0.158	
		15	5	1	0.086	0.061	0.043	0.040	0.027	0.095
				2	0.170	0.090	0.054	0.067	0.029	0.096
				3	0.203	0.093	0.051	0.087	0.033	0.094

Table 3 Average bias values obtained from the parameter recovery experiment

J	I	R	L	θ_{ji}	α_{ii}	α_r	β_i	β_r	d_{im}		
50	5	5	1	0.001	-0.025	0.006	-0.007	0.000	0.000		
			2	0.005	-0.102	0.005	0.002	0.000	0.000		
			3	-0.001	0.156	-0.012	-0.012	0.000	0.000		
	15	5	1	-0.003	0.026	-0.011	-0.004	0.000	0.000		
			2	0.001	-0.088	0.008	0.003	0.000	0.000		
			3	-0.002	-0.141	-0.017	0.011	0.000	0.000		
	15	5	5	1	0.007	0.019	0.001	-0.004	0.000	0.000	
				2	0.005	0.023	0.005	-0.005	0.000	0.000	
				3	-0.002	-0.052	-0.003	-0.001	0.000	0.000	
		15	5	5	1	0.009	-0.010	-0.002	0.010	0.000	0.000
					2	0.001	-0.044	-0.002	-0.001	0.000	0.000
					3	0.000	-0.038	-0.005	-0.009	0.000	0.000
		100	5	5	1	0.001	0.020	-0.008	-0.004	0.000	0.000
					2	0.001	0.087	0.005	0.005	0.000	0.000
					3	0.001	-0.103	-0.009	0.000	0.000	0.000
15	5		5	1	-0.001	0.001	-0.006	0.001	0.000	0.000	
				2	-0.001	-0.040	-0.004	-0.002	0.000	0.000	
				3	-0.001	0.072	0.005	0.000	0.000	0.000	
15	5		5	1	-0.005	-0.022	-0.003	0.001	0.000	0.000	
				2	0.003	0.022	0.001	0.006	0.000	0.000	
				3	-0.002	-0.036	0.002	-0.005	0.000	0.000	
	15	5	5	1	0.002	0.006	-0.001	0.002	0.000	0.000	
				2	0.002	-0.025	-0.001	0.000	0.000	0.000	
				3	-0.001	-0.023	-0.002	0.000	0.000	0.000	

0.000 indicates that the absolute value of the average bias was less than 0.001

7.2 Validity of dimensionality selection using information criteria

This subsection describes a simulation experiment for evaluating the accuracy of the dimensionality selection using the WAIC and WBIC as information criteria. Concretely, we conducted the following experiments by changing the number of examinees, evaluation items, raters and dimensions to $J \in \{50, 100\}$, $I \in \{5, 15\}$, $R \in \{5, 15\}$, and $L \in \{1, 2, 3\}$, respectively.

1. For J examinees, I evaluation items, R raters, and L dimensions, generate rating data from the proposed model after the true model parameters are randomly generated from the distributions in Eq. (8). The number of categories, K , was fixed to 4, as in the parameter recovery experiment.
2. Using the generated data, estimate the parameters in the proposed model and calculate the WAIC and the WBIC values while changing the number of dimensions L^e to $\{1, 2, 3\}$.

3. Rank the WAIC and WBIC values for each L^e , such that the L^e with the lowest WAIC and WBIC values is ranked first.
4. Repeat the above procedure 30 times, then calculate the average rank. Additionally, calculate the ratio of the correct dimensionality identification for each setting.

Table 4 shows the results. The $L^e = 1, 2, 3$ columns show the average of the estimated ranks, with the highest average rank for each setting shown in bold. The *Acc* column shows the ratio of the correct dimensionality identification.

The results show that the WAIC can select true dimensionality in many cases. The WBIC can also select true dimensionality when the size of the data increases, although it tends to be inferior to the WAIC when the size of the data is small.

Table 4 Results of model comparison using an information criterion

<i>J</i>	<i>I</i>	<i>R</i>	<i>L</i>	WAIC				WBIC				
				$L^e = 1$	$L^e = 2$	$L^e = 3$	Acc	$L^e = 1$	$L^e = 2$	$L^e = 3$	Acc	
50	5	5	1	1.533	1.733	2.733	0.667	3.000	1.970	1.030	0.000	
			2	2.800	1.633	1.567	0.367	2.930	1.900	1.170	0.133	
			3	3.000	1.833	1.167	0.833	2.770	1.930	1.300	0.800	
	15	1	1	1.333	1.800	2.867	0.767	1.300	2.030	2.670	0.800	
			2	3.000	1.300	1.700	0.700	1.970	1.930	2.100	0.133	
			3	3.000	1.967	1.033	0.967	2.430	2.000	1.570	0.700	
	15	5	1	1	1.233	2.033	2.733	0.800	1.000	2.000	3.000	1.000
				2	3.000	1.200	1.800	0.800	1.470	1.730	2.800	0.267
				3	3.000	2.000	1.000	1.000	1.400	1.970	2.630	0.133
15		1	1	1.133	1.967	2.900	0.900	1.000	2.000	3.000	1.000	
			2	3.000	1.067	1.933	0.933	2.970	1.000	2.030	1.000	
			3	3.000	2.000	1.000	1.000	3.000	1.670	1.330	0.667	
100	5	5	1	1.567	2.067	2.367	0.700	3.000	2.000	1.000	0.000	
			2	2.933	1.767	1.300	0.267	3.000	1.970	1.030	0.033	
			3	3.000	2.000	1.000	1.000	2.900	2.070	1.030	0.967	
		15	1	1	1.300	1.933	2.767	0.800	1.400	2.030	2.570	0.733
				2	3.000	1.333	1.667	0.667	2.200	1.970	1.830	0.167
				3	3.000	1.933	1.067	0.933	2.330	2.100	1.570	0.667
	15	5	1	1	1.200	1.900	2.900	0.833	1.000	2.000	3.000	1.000
				2	3.000	1.200	1.800	0.800	1.070	2.000	2.930	0.033
				3	3.000	2.000	1.000	1.000	1.270	2.000	2.730	0.133
		15	1	1	1.167	1.867	2.967	0.867	1.000	2.000	3.000	1.000
				2	3.000	1.133	1.867	0.867	2.730	1.130	2.130	0.867
				3	3.000	2.000	1.000	1.000	3.000	1.670	1.330	0.667

These results suggest that both the information criteria can find the optimal number of dimensions for larger-scale settings, although the WAIC will be more accurate than the WBIC in relatively small-scale settings.

7.3 Accuracy of ability measurement

This subsection evaluates whether the consideration of the rater characteristics in the proposed model is effective in improving ability measurement accuracy. For the evaluation, we compare ability measurement accuracy between the proposed model and the conventional multidimensional GPCM. Note that the conventional multidimensional GPCM is consistent with the proposed model without rater parameters. We conducted the following experiments by changing the number of examinees, evaluation items, raters, and dimensions to $J \in \{50, 100\}$, $I \in \{5, 15\}$, $R \in \{5, 15\}$, and $L \in \{1, 2, 3\}$, respectively.

1. For J examinees, I evaluation items, R raters, and L dimensions, generate true model parameters randomly from the proposed model and the conventional multidimensional GPCM. Here, the parameters were generated from the distributions in Eq. (8). The number of categories, K , was fixed to 4, as in the above experiments.
2. Generate rating data from the two models respectively given the parameters generated in Procedure 1.
3. Using each generated dataset, estimate the model parameters and examinee abilities in each model.
4. Calculate Pearson's correlation between the ability estimates and the true ability values generated in Procedure 1.
5. Repeat the above procedure 30 times, then calculate the average correlation value.

Note that the conventional multidimensional GPCM directly cannot handle three-way data consisting of examinees, raters, and evaluation items. Thus, in this experiment, we applied the conventional multidimensional GPCM assuming that the probability for the three-way data P_{ijk} is defined by P_{ijk} shown in Eq. (5) as follows:

$$P_{ijk} = P_{ijk} = \frac{\exp \sum_{m=1}^k \left[1.7 \left(\sum_{l=1}^L \alpha_{il} \theta_{jl} - \beta_i - d_{im} \right) \right]}{\sum_{l=1}^K \exp \sum_{m=1}^l \left[1.7 \left(\sum_{l=1}^L \alpha_{il} \theta_{jl} - \beta_i - d_{im} \right) \right]}. \quad (10)$$

Table 5 shows the results. In the table, the Generation from Prop and Generation from Conv columns show the results of data are generated from the proposed model and the conventional multidimensional GPCM, respectively. The sub-columns Prop and Conv show the average correlation values when the proposed model and the conventional multidimensional GPCM are applied to each dataset respectively. Furthermore, the sub-column Diff indicates the difference in average correlation values between Prop and Conv, where a larger Diff value

Table 5 Ability measurement accuracy for complete data under a fully crossed design

<i>J</i>	<i>I</i>	<i>R</i>	<i>L</i>	Generation from Prop				Generation from Conv			
				Prop	Conv	Diff	<i>p</i>	Prop	Conv	Diff	<i>p</i>
50	5	5	1	0.980	0.973	0.007	< 0.01	0.979	0.979	0.000	0.12
			2	0.954	0.944	0.010	< 0.01	0.947	0.947	0.000	0.10
			3	0.935	0.891	0.044	< 0.05	0.926	0.926	0.000	0.35
		15	1	0.994	0.990	0.004	< 0.01	0.993	0.993	0.000	< 0.05
			2	0.976	0.967	0.009	< 0.01	0.971	0.971	0.000	0.27
			3	0.962	0.946	0.016	< 0.01	0.957	0.958	−0.001	< 0.01
	15	5	1	0.994	0.990	0.004	< 0.01	0.993	0.993	0.000	< 0.01
			2	0.969	0.959	0.010	< 0.01	0.967	0.967	0.000	0.62
			3	0.957	0.938	0.019	< 0.01	0.948	0.948	0.000	0.22
		15	1	0.998	0.996	0.002	< 0.01	0.998	0.998	0.000	0.46
			2	0.985	0.975	0.010	< 0.01	0.981	0.981	0.000	0.47
			3	0.981	0.965	0.016	< 0.01	0.975	0.975	0.000	0.23
100	5	5	1	0.981	0.975	0.006	< 0.01	0.978	0.978	0.000	0.34
			2	0.959	0.948	0.011	< 0.01	0.951	0.952	0.000	0.14
			3	0.941	0.894	0.047	< 0.05	0.931	0.932	0.000	< 0.05
		15	1	0.994	0.989	0.005	< 0.01	0.992	0.993	0.000	< 0.01
			2	0.983	0.974	0.009	< 0.01	0.976	0.976	0.000	0.96
			3	0.972	0.960	0.012	< 0.01	0.963	0.964	0.000	0.54
	15	5	1	0.994	0.991	0.003	< 0.01	0.993	0.993	0.000	0.98
			2	0.975	0.968	0.007	< 0.01	0.973	0.973	0.000	0.94
			3	0.967	0.949	0.018	< 0.01	0.960	0.960	0.000	0.28
		15	1	0.998	0.996	0.002	< 0.01	0.998	0.998	0.000	0.61
			2	0.989	0.982	0.007	< 0.01	0.988	0.988	0.000	0.16
			3	0.983	0.973	0.010	< 0.01	0.979	0.979	0.000	0.40
Avg.				0.976	0.964	0.012	–	0.971	0.972	0.000	–

indicates that the proposed model is more accurate. We also conducted the paired *t*-test for the averaged agreement correlation between the proposed model and the conventional multidimensional GPCM, and the resulting *p* values are shown in *p* column.

The results show that the performance of the conventional model significantly drops when the data are generated from the proposed model, whereas the performance of the proposed model is almost equal to that of the conventional model when the data are generated from the conventional model. This suggests that lack of knowledge of the rater characteristics deteriorates ability measurement accuracy when the raters are assumed to have different characteristics. Note that the *p* values in the paired *t*-test depend on two factors, namely, (1) the mean of the difference between conditions, and (2) the standard deviation of the difference between conditions. Thus, as seen in Table 5, a larger absolute Diff value does not necessarily result in a lower *p* value.

Table 6 Example of fully crossed design

	Rater				
	1	2	3	4	5
Examinee 1	✓	✓	✓	✓	✓
Examinee 2	✓	✓	✓	✓	✓
Examinee 3	✓	✓	✓	✓	✓
Examinee 4	✓	✓	✓	✓	✓
Examinee 5	✓	✓	✓	✓	✓

Table 7 Example of systematic link design

	Rater				
	1	2	3	4	5
Examinee 1	✓	✓			
Examinee 2		✓	✓		
Examinee 3			✓	✓	
Examinee 4				✓	✓
Examinee 5	✓				✓

Although the results demonstrate the effectiveness of the proposed model, Table 5 shows that improvement in correlation values by the proposed model is small. This is because data X are generated as complete data under a fully crossed design, assuming all the raters evaluate all the examinees. In this case, because the data per examinee are large and dense, ability measurement accuracy tends to be extremely high in both the models, making the difference in performance among the models small. However, in practice, the data will be sparser because we often assign few raters for each examinee to decrease the raters' assessment workload. In sparse data settings, we can expect the difference in performance among the models to be clearer.

Thus, we conducted the same experiment as described above assuming a practice situation where few raters are assigned to each examinee. Concretely, in Procedure 2, we first assigned two raters to each examinee based on a systematic link design (Shin et al. 2019; Uto 2020; Wind and Jones 2019), and then we generated the data based on the rater assignment. The examples of a fully crossed design and a systematic link design are illustrated in Tables 6 and 7, where checkmarks indicate an assigned rater, and blank cells indicate that no rater was assigned. The data without assigned raters are treated as missing data.

Table 8 shows the results. The results show that the average correlation values of the conventional model drops substantially when the data are generated from the proposed model, whereas the high performance of the proposed model is still maintained, regardless of data generation models.

Table 8 Ability measurement accuracy for sparse data under a systematic links design

<i>J</i>	<i>I</i>	<i>R</i>	<i>L</i>	Generation from Prop				Generation from Conv			
				Prop	Conv	Diff	<i>p</i>	Prop	Conv	Diff	<i>p</i>
50	5	5	1	0.937	0.837	0.100	< 0.01	0.948	0.952	−0.004	< 0.01
			2	0.898	0.795	0.103	< 0.01	0.897	0.904	−0.006	0.29
			3	0.844	0.653	0.191	< 0.01	0.874	0.877	−0.003	< 0.05
		15	1	0.907	0.770	0.137	< 0.01	0.929	0.946	−0.018	< 0.01
			2	0.870	0.723	0.148	< 0.01	0.897	0.908	−0.011	< 0.01
			3	0.828	0.642	0.185	< 0.01	0.850	0.859	−0.008	0.07
	15	5	1	0.979	0.866	0.113	< 0.01	0.981	0.982	−0.001	< 0.05
			2	0.932	0.853	0.080	< 0.01	0.937	0.938	−0.001	< 0.05
			3	0.906	0.799	0.107	< 0.01	0.914	0.911	0.003	0.40
		15	1	0.972	0.833	0.139	< 0.01	0.977	0.983	−0.006	< 0.01
			2	0.929	0.856	0.073	< 0.01	0.937	0.942	−0.005	< 0.01
			3	0.900	0.740	0.160	< 0.01	0.910	0.914	−0.004	< 0.01
100	5	5	1	0.940	0.800	0.139	< 0.01	0.946	0.949	−0.003	< 0.01
			2	0.903	0.843	0.060	< 0.01	0.910	0.911	−0.001	< 0.01
			3	0.879	0.686	0.193	< 0.01	0.882	0.883	−0.001	< 0.01
		15	1	0.931	0.778	0.153	< 0.01	0.935	0.946	−0.010	< 0.01
			2	0.892	0.800	0.093	< 0.01	0.893	0.900	−0.006	< 0.01
			3	0.865	0.724	0.141	< 0.01	0.876	0.882	−0.005	< 0.01
	15	5	1	0.977	0.856	0.121	< 0.01	0.982	0.983	−0.001	< 0.01
			2	0.938	0.858	0.081	< 0.01	0.947	0.948	0.000	0.10
			3	0.917	0.840	0.077	< 0.01	0.921	0.921	0.000	< 0.05
		15	1	0.974	0.819	0.155	< 0.01	0.979	0.982	−0.003	< 0.01
			2	0.935	0.854	0.081	< 0.01	0.945	0.947	−0.002	< 0.01
			3	0.915	0.860	0.055	< 0.01	0.920	0.921	−0.001	< 0.01
Avg.			0.915	0.795	0.120	–	0.925	0.929	−0.004	–	

From these experimental results, we can conclude that the consideration of the rater characteristics in the proposed model is effective in improving ability measurement accuracy.

8 Actual data experiments

This section describes the performance of the proposed model in experiments based on actual data.

8.1 Actual data

In this experiment, actual rubric-based performance assessment data were gathered as follows:

1. We recruited 134 Japanese university students as participants.
2. The participants were asked to complete an essay-writing task that involved translating a task taken from the National Assessment of Educational Progress assessments (Persky et al. 2003; Salah-Din et al. 2008). No specific or preliminary knowledge was needed to complete the task.
3. The written essays were evaluated by 18 raters using a rubric consisting of 9 evaluation items divided into 4 rating categories. We assigned four raters to each essay based on a systematic links design (Shin et al. 2019; Uto 2020; Wind and Jones 2019) to reduce the raters' assessment workload. The evaluation items column in Table 9 lists the abstracts of the evaluation items in the rubric, and was created based on two writing assessment rubrics proposed by Matsushita et al. (2013), Nakajima (2017) for Japanese university students. Furthermore, Appendix 2 presents all the information in the rubric.

We evaluated the effectiveness of the proposed model using the obtained data.

8.2 Model comparison using information criteria

As explained above, the proposed model can estimate examinee ability on a multidimensional scale while considering the characteristics of both the raters and the rubric's evaluation items. To evaluate the effectiveness of the consideration of the multidimensionality and rater characteristics, we conducted a model fitting

Table 9 Evaluation items and their characteristic parameters

Evaluation items	$\hat{\alpha}_{i1}$	$\hat{\alpha}_{i2}$	$\hat{\beta}_i$	\hat{d}_{i2}	\hat{d}_{i3}	\hat{d}_{i4}
1 Appropriateness of problem setting	0.203	0.381	-0.631	-0.920	-0.250	1.170
2 Consistency between claims and conclusions	0.222	0.473	-0.629	-0.731	-0.291	1.022
3 Presentation of evidence	0.137	0.451	-0.595	-1.472	0.219	1.254
4 Consideration of opposing viewpoints	0.111	0.274	-0.297	-0.399	-0.246	0.645
5 Appropriateness of logical structure	0.296	0.495	-0.795	-0.901	-0.212	1.113
6 Consideration of readers	0.314	0.442	-0.673	-0.651	-0.460	1.112
7 Typographical accuracy	0.517	0.109	-1.345	-0.814	-0.380	1.194
8 Stylistic consistency	0.421	0.177	-0.815	-0.659	-0.247	0.907
9 Usage of conjunctions	0.449	0.243	-1.250	-0.762	-0.282	1.044

Table 10 Model comparison using actual data

	L				
	1	2	3	4	5
WAIC					
Proposed	11742.00	11647.73	11654.13	11654.73	11660.79
Conventional	12280.39	12200.08	12207.81	12220.04	12229.58
WBIC					
Proposed	6395.02	6390.03	6402.01	6428.20	6449.91
Conventional	6584.02	6570.78	6581.95	6598.59	6630.79

evaluation based on information criteria. Specifically, we calculated the WAIC and WBIC for the proposed model and the conventional multidimensional GPCM, consistent with the proposed model without rater parameters, using the actual data for each dimensionality $L \in \{1, \dots, 5\}$.

Table 10 shows the results, with the minimum score for each criteria in bold. The table indicates that the WAIC and WBIC are minimized when $L = 2$ in both the proposed model and the conventional model. This means that the unidimensionality assumption is not satisfied in the data, suggesting the requirement of the multidimensional models. Furthermore, comparison of the two models shows that the proposed model provides better model fitting than the conventional model in all cases. The results suggest that consideration of rater characteristics is effective in improving model fitting, which verifies the effectiveness of the proposed model.

8.3 Characteristic interpretation of the rubric's evaluation items

In this subsection, we show the interpretation of the characteristics of the evaluation items. Table 9 shows the parameters of the evaluation items, which were estimated by the proposed model under $L = 2$. Here, $L = 2$ was used because it provided the best model fitting, as shown in the experiment above.

According to Table 9, the evaluation items reveal different patterns of discrimination parameters. For example, evaluation items 1–6 have larger discrimination values in the second dimension, whereas evaluation items 7–9 have larger discrimination values in the first dimension. Moreover, evaluation item 4 has relatively low discrimination values in both dimensions, meaning that it might not be suitable for distinguishing examinee ability. In contrast, evaluation item 6 has moderate discrimination values in both dimensions, meaning that it measures two-dimensional ability concurrently.

The discrimination parameters of each evaluation item enable us to interpret what is mainly measured by each ability dimension. Specifically, as described above,

Table 9 shows that evaluation items 1–6 have larger discrimination values in the second dimension, and evaluation items 7–9 have larger values in the first dimension. These results suggest that the first ability dimension reflects a common ability underlying evaluation items 7–9, and the second dimension reflects a common ability underlying evaluation items 1–6. According to the contents of the evaluation items (see Appendix 2), we can see that evaluation items 7–9 relate to stylistic skills (such as typological errors and word choice), whereas evaluation items 1–6 relate to logical skills (such as augmentation and organization). Indeed, the rubric was designed such that evaluation items 1–6 mainly measure argumentative skills, and evaluation items 7–9 measure stylistic skills (Matsushita et al. 2013; Nakajima 2017). These results suggest that the rubric developer’s expectation is supported by the analysis based on the proposed model.

Furthermore, Table 9 shows that the level of difficulty differs among the evaluation items. For example, evaluation item 4 is the most difficult, and evaluation item 7 is the easiest. These are reasonable judgments because evaluation item 4 requires sufficient discussion about opposing opinions, whereas evaluation item 7 requires only superficial typological correctness.

The step difficulty parameters, d_{im} , also show different patterns, meaning that the score distribution differs among the evaluation items. As examples, Fig. 4 depicts the IRSs for evaluation items 3 and 4, which have different step difficulty parameter patterns as well as relatively similar discrimination and difficulty. The figure shows that a score of 2 tends to be avoided and a score of 3 tends to be preferred in evaluation item 4.

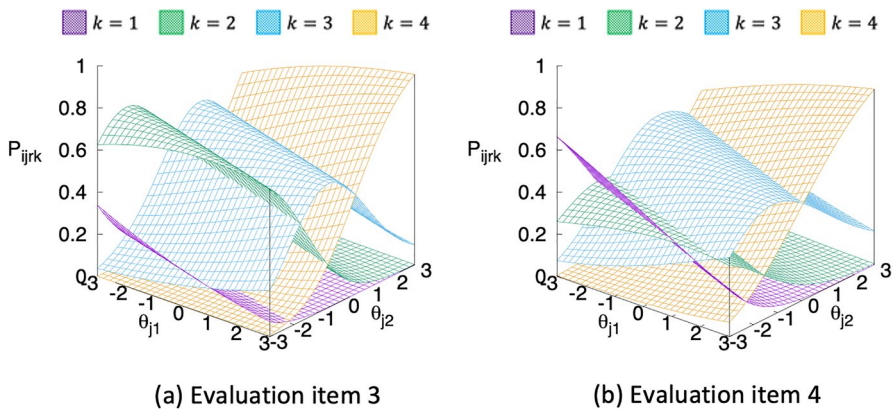
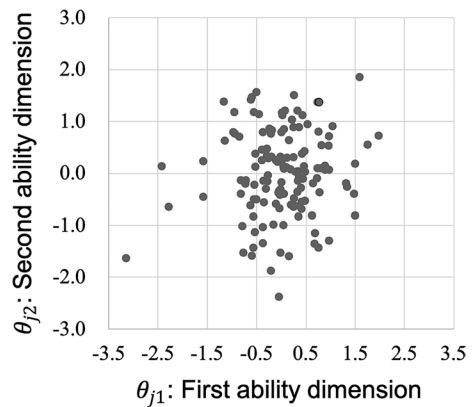


Fig. 4 Item response surfaces for two evaluation items with different step difficulty

Table 11 Rater parameter estimates

r	$\hat{\alpha}_r$	$\hat{\beta}_r$	r	$\hat{\alpha}_r$	$\hat{\beta}_r$
1	0.579	0.554	10	1.001	-0.038
2	1.007	0.137	11	0.771	0.032
3	0.675	-0.447	12	1.761	0.093
4	1.232	0.593	13	0.603	0.242
5	1.368	0.196	14	1.051	-0.362
6	1.881	0.200	15	1.845	-0.204
7	0.827	-0.195	16	1.169	0.143
8	1.692	0.545	17	0.450	-0.458
9	0.657	-1.075	18	1.044	0.044

Fig. 5 Ability estimates

8.4 Rater parameter estimates and ability estimates

To confirm whether the rater characteristics differed, rater parameter estimates were obtained, as shown in Table 11. According to the table, severity and consistency differ among raters. For example, Raters 1 and 17 are highly inconsistent raters whose ratings might be unreliable, whereas Raters 8 and 15 are highly consistent raters. Furthermore, Raters 1, 4, and 8 have higher severity values, whereas Raters 3, 9 and 17 have lower severity values. The variety of rater characteristics is the reason why the proposed model provided better model fitting than the conventional multidimensional GPCM.

Moreover, Fig. 5 shows the two-dimensional ability estimates for each examinee. The horizontal axis indicates the first-dimensional ability value θ_{j1} , the vertical axis indicates the second-dimensional ability value θ_{j2} , and each dot represents an examinee. The figure shows that the examinees have different ability

patterns. Such multidimensional ability measurement cannot be realized by conventional unidimensional IRT models.

9 Conclusion

This study proposed a new IRT model for rubric-based performance assessment. The model was formulated as a multidimensional extension of the generalized MFRM. A NUT variant of the HMC algorithm for the proposed model was implemented using the software package Stan. Through simulation experiments, we demonstrated the following: (1) The MCMC algorithm appropriately estimates the model parameters. (2) An optimal number of dimensions for the proposed model can be determined using information criteria. (3) The consideration of the rater characteristics in the proposed model is effective in improving ability measurement accuracy. We also conducted real data application experiments to show examples of analysis of rubric quality and rubric construct validity by interpreting the dimensionality and the characteristics of the evaluation items. Also, the actual data experiment showed that the consideration of the multidimensionality and rater characteristics in the proposed model improved the model fitting.

In future studies, we plan to evaluate the effectiveness of the proposed model using various and more massive datasets. Furthermore, we hope to extend the proposed model to four-way data consisting of examinees \times raters \times evaluation items \times performance tasks because practical tests often include several tasks.

Appendix 1

The Stan code for the proposed model is as follows:

```

data{
  int <lower=0> J; // # of examinees
  int <lower=0> I; // # of evaluation items
  int <lower=0> R; // # of raters
  int <lower=2> K; // # of rating categories
  int <lower=0> D; // # of dimensions
  int <lower=0> N; // # of rating data
  int <lower=1, upper=J> ExamineeID [N]; // list of examinee ID
  int <lower=1, upper=I> ItemID [N]; // list of evaluation item ID
  int <lower=1, upper=R> RaterID [N]; // list of rater ID
  int <lower=1, upper=K> X [N]; // list of ratings
}
transformed data{
  vector[K] c = cumulative_sum(rep_vector(1, K)) - 1;
}
parameters {
  vector[D] theta[J];
  real<lower=0> alpha_r [R-1];
  vector<lower=0>[D] alpha_i [I];
  vector[R-1] beta_r;
  vector[I] beta_i;
  vector[K-2] beta_ik [I];
}
transformed parameters{
  real<lower=0> trans_alpha_r[R];
  vector[R] trans_beta_r;
  vector[K-1] category_est[I];
  vector[K] category_prm[I];
  trans_alpha_r[1] = 1.0 / prod(alpha_r);
  trans_beta_r[1] = -1*sum(beta_r);
  trans_alpha_r[2:R] = alpha_r;
  trans_beta_r[2:R] = beta_r;
  for(z in 1:I){
    category_est[z, 1:(K-2)] = beta_ik [z];
    category_est[z, K-1] = -1*sum(beta_ik [z]);
    category_prm[z] = cumulative_sum(append_row(0, category_est[z]));
  }
}
model{
  for (d in 1:D){
    theta[d,] ~ normal(0, 1);
    alpha_i[d,] ~ lognormal(0, 1.0);
  }
  trans_alpha_r ~ lognormal(0, 1.0);
  trans_beta_r ~ normal(0, 1);
  beta_i ~ normal(0, 1);
  for (z in 1:I) category_est [z,] ~ normal(0, 1);
  for (n in 1:N){
    X[n] ~ categorical_logit(1.7 * trans_alpha_r[RaterID[n]] * (c*(dot_product
      (alpha_i[ItemID[n]], theta[ExamineeID[n]])-beta_i[ItemID[n]]-
      trans_beta_r[RaterID[n]]-category_prm[ItemID[n]]));
  }
}
generated quantities {
  vector[N] log_lik;
  for (n in 1:N){
    log_lik[n] = categorical_logit_log(X[n], 1.7 * trans_alpha_r[RaterID[n]] * (
      c*(dot_product(alpha_i[ItemID[n]], theta[ExamineeID[n]])-beta_i[
      ItemID[n]]-trans_beta_r[RaterID[n]]-category_prm[ItemID[n]]));
  }
}

```

Appendix 2

The rubric used in this study is shown in Tables 12 and 13.

Table 12 Evaluation items 1, 2, 3, and 4 in the rubric

Evaluation items	Rating categories			
	4	3	2	1
Appropriateness of problem setting	The problem described adheres to the given theme, and descriptions of why the problem was addressed and its background include an explanation of its importance	The problem described adheres to the given theme, and there is a description of why the problem was addressed and its background	The problem described adheres to the given theme, but there is insufficient description of why the problem was addressed or its background	Does not meet criteria for a score of 2
Consistency between claims and conclusions	Derives a conclusion corresponding to the author's assertions about the problem described. The conclusion is novel and goes beyond general theory	Derives a conclusion corresponding to the author's assertions about the problem described	A conclusion is described, but it insufficiently corresponds to the author's assertions	Does not meet criteria for a score of 2
Presentation of evidence	Evidence for the author's assertions is presented with the support of various reliable facts and data	Evidence for the author's assertions is presented with the support of reliable facts or data in at least one case	Evidence for the author's assertions is presented, but without the support of reliable facts or data in at least one case	Does not meet criteria for a score of 2
Consideration of opposing viewpoints	Opposing viewpoints to the author's assertions are presented, along with refutations (indications of problems) to those viewpoints	At least one opposing viewpoint to the author's assertions is presented, along with a refutation to that viewpoint	Opposing viewpoints to the author's assertions are presented, but with no refutations to those viewpoints	Does not meet criteria for a score of 2

Table 13 Evaluation items 5, 6, 7, 8, and 9 in the rubric

Evaluation items	Rating categories			
	4	3	2	1
Appropriateness of logical structure	From description of the problem to the conclusion, there are highly logical structures, orderly descriptions, and connectivity between paragraphs	From description of the problem to the conclusion, there are somewhat logical structures, orderly descriptions, and connectivity between paragraphs	There is a flow from description of the problem to the conclusion, but there are some problems in orderly descriptions and connectivity between paragraphs	Does not meet criteria for a score of 2
Consideration of readers	Specific descriptions are presented, demonstrating awareness of the flow of argument and consideration of the readers' thought processes	Specific descriptions are presented, demonstrating consideration of the readers' thought processes, but there is room for improvement	Attempts are made toward considering the readers' thought processes, but there are many deficits, such as lack of specificity	Does not meet criteria for a score of 2
Typographical accuracy	There are no typographical errors	There are 1–3 typographical errors	There are 4–6 typographical errors	Does not meet criteria for a score of 2
Stylistic consistency	There is stylistic consistency and appropriate use of written language, which consists of easy-to-understand sentences and phrases	The author demonstrates efforts toward stylistic consistency and appropriate use of written language, but there are 1–3 mistakes	There is general stylistic consistency, but also 4 or more mistakes such as use of spoken language or ambiguous meanings	Does not meet criteria for a score of 2
Usage of conjunctions	Connective phrases are appropriately used	There are 1–2 misuses of connective phrases	There are 3–5 misuses of connective phrases	Does not meet criteria for a score of 2

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers 19H05663 and 21H00898.

Declarations

Conflict of interest The authors have no conflict of interest directly relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* 19:716–723
- Almond RG (2014) A comparison of two MCMC algorithms for hierarchical mixture models. In: *Proceedings of the uncertainty in artificial intelligence conference on Bayesian modeling applications workshop*, pp 1–19
- Andrich D (1978) A rating formulation for ordered response categories. *Psychometrika* 43(4):561–573
- Baker F, Kim SH (2004) *Item response theory: parameter estimation techniques*. Marcel Dekker, New York
- Brooks S, Gelman A, Jones G, Meng X (2011) *Handbook of Markov chain Monte Carlo*. CRC Press, Boca Raton
- Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M et al (2017) Stan: a probabilistic programming language. *J Stat Softw* 76(1):1–32
- Chan S, Bax S, Weir C (2017) Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors (Tech. Rep.). *IELTS Research Reports Online Series*
- DeCarlo LT (2005) A model of rater behavior in essay grading based on signal detection theory. *J Educ Meas* 42(1):53–76
- DeCarlo LT, Kim YK, Johnson MS (2011) A hierarchical rater model for constructed responses, with a signal detection rater model. *J Educ Meas* 48(3):333–356
- Deng S, McCarthy DE, Piper ME, Baker TB, Bolt DM (2018) Extreme response style and the measurement of intra-individual variability in affect. *Multivar Behav Res* 53(2):199–218
- Eckes T (2015) *Introduction to many-facet Rasch measurement: analyzing and evaluating rater-mediated assessments*. Peter Lang Pub. Inc, New York
- Elliott M, Haviland A, Kanouse D, Hambarsoomian K, Hays R (2009) Adjusting for subgroup differences in extreme response tendency in ratings of health care: impact on disparity estimates. *Health Serv Res* 44:542–561
- Fontanella L, Fontanella S, Valentini P, Trendafilov N (2019) Simple structure detection through Bayesian exploratory multidimensional IRT models. *Multivar Behav Res* 54(1):100–112
- Fox J-P (2010) *Bayesian item response modeling: theory and applications*. Springer, Berlin
- Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Stat Sci* 7(4):457–472
- Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2013) *Bayesian data analysis*, 3rd edn. Taylor & Francis, New York
- Girolami M, Calderhead B (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J R Stat Soc Ser B Stat Methodol* 73(2):123–214

- Hoffman MD, Gelman A (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 15:1593–1623
- Hua C, Wind SA (2019) Exploring the psychometric properties of the mind-map scoring rubric. *Behaviormetrika* 46(1):73–99
- Hussein MA, Hassan HA, Nassef M (2019) Automated language essay scoring systems: a literature review. *PeerJ Comput Sci* 5:e208
- Hutten LR (1980) Some empirical evidence for latent trait model selection. ERIC Clearinghouse, Washington
- Jiang Z, Carter R (2019) Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behav Res Methods* 51(2):651–662
- Jin K-Y, Wang W-C (2018) A new facets model for rater's centrality/extremity response style. *J Educ Meas* 55(4):543–563
- Kaliski PK, Wind SA, Engelhard G, Morgan DL, Plake BS, Reshetar RA (2013) Using the many-faceted Rasch model to evaluate standard setting judgments. *Educ Psychol Meas* 73(3):386–411
- Kose IA, Demirtasli NC (2012) Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Proc Soc Behav Sci* 46:135–140
- Linacre JM (1989) Many-faceted Rasch measurement. MESA Press, San Diego
- Linlin C (2019) Comparison of automatic and expert teachers' rating of computerized English listening-speaking test. *Engl Lang Teach* 13(1):18
- Lord F (1980) Applications of item response theory to practical testing problems. Erlbaum Associates, Mahwah
- Louvigné S, Uto M, Kato Y, Ishii T (2018) Social constructivist approach of motivation: social media messages recommendation system. *Behaviormetrika* 45(1):133–155
- Luo Y, Al-Harbi K (2017) Performances of LOO and WAIC as IRT model selection methods. *Psychol Test Assess Model* 59(2):183–205
- Luo Y, Jiao H (2018) Using the Stan program for Bayesian item response theory. *Educ Psychol Meas* 78(3):384–408
- Martin-Fernandez M, Revuelta J (2017) Bayesian estimation of multidimensional item response models. A comparison of analytic and simulation algorithms. *Int J Methodol Exp Psychol* 38(1):25–55
- Masters G (1982) A Rasch model for partial credit scoring. *Psychometrika* 47(2):149–174
- Matsushita K, Ono K, Takahashi Y (2013) Development of a rubric for writing assessment and examination of its reliability. *J Lib Gen Educ Soc Jpn* 35(1):107–115 (in Japanese)
- Mislevy RJ (2018) Sociocognitive foundations of educational measurement. Routledge, London
- Muraki E (1997) A generalized partial credit model. In: van der Linden WJ, Hambleton RK (eds) *Handbook of modern item response theory*. Springer, Berlin, pp 153–164
- Murtonen M, Balloo K (2019) Redefining scientific thinking for higher education: higher-order thinking, evidence-based reasoning and research skills. Palgrave Macmillan, London
- Myford CM, Wolfe EW (2003) Detecting and measuring rater effects using many-facet Rasch measurement: part I. *J Appl Meas* 4:386–422
- Myford CM, Wolfe EW (2004) Detecting and measuring rater effects using many-facet Rasch measurement: part II. *J Appl Meas* 5:189–227
- Nakajima A (2017) Achievements and issues in the application of rubrics in academic writing: a case study of the college of images arts and sciences. *Ritsumeikan High Educ Stud* 17:199–215 (in Japanese)
- Nguyen T, Uto M, Abe Y, Ueno M (2015) Reliable peer assessment for team project based learning using item response theory. In: *Proceedings of the international conference on computers in education*, pp 144–153
- Palm T (2008) Performance assessment and authentic assessment: a conceptual analysis of the literature. *Pract Assess Res Eval* 13(4):1–11
- Patz RJ, Junker B (1999) Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *J Educ Behav Stat* 24(4):342–366
- Patz RJ, Junker BW, Johnson MS, Mariano LT (2002) The hierarchical rater model for rated test items and its application to largescale educational assessment data. *J Educ Behav Stat* 27(4):341–384
- Persky H, Daane M, Jin Y (2003) The nation's report card: writing 2002 (Tech. Rep.). National Center for Education Statistics

- Rahman AA, Hanafi NM, Yusof Y, Mukhtar MI, Yusof AM, Awang H (2020) The effect of rubric on rater's severity and bias in TVET laboratory practice assessment: analysis using many-facet Rasch measurement. *J Tech Educ Train* 12(1):57–67
- Reckase MD (2009) *Multidimensional item response theory models*. Springer, Berlin
- Reise SP, Revicki DA (2014) *Handbook of item response theory modeling: applications to typical performance assessment*. Routledge, London
- Salahu-Din D, Persky H, Miller J (2008) *The nation's report card: writing 2007 (Tech. Rep.)*. National Center for Education Statistics
- Schwarz G (1978) Estimating the dimensions of a model. *Ann Stat* 6:461–464
- Shavelson RJ, Zlatkin-Troitschanskaia O, Beck K, Schmidt S, Marino JP (2019) Assessment of university students' critical thinking: next generation performance assessment. *Int J Test* 19(4):337–362
- Shin HJ, Rabe-Hesketh S, Wilson M (2019) Trifactor models for multiple-ratings data. *Multivar Behav Res* 54(3):360–381
- Soo Park Y, Xing K (2019) Rater model using signal detection theory for latent differential rater functioning. *Multivar Behav Res* 54(4):492–504
- Stan Development Team (2018) RStan: the R interface to stan. R package version 2.17.3. <http://mc-stan.org>
- Svetina D, Valdivia A, Underhill S, Dai S, Wang X (2017) Parameter recovery in multidimensional item response theory models under complexity and nonnormality. *Appl Psychol Meas* 41(7):530–544
- Tavakol M, Pinner G (2019) Using the many-facet Rasch model to analyse and evaluate the quality of objective structured clinical examination: a non-experimental cross-sectional design. *BMJ Open* 9(9):1–9
- Uto M (2019) Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In: *Proceedings of the international conference on artificial intelligence in education*, pp 494–506
- Uto M (2020) Accuracy of performance-test linking based on a many-facet Rasch model. *Behav Res Methods*. <https://doi.org/10.3758/s13428-020-01498-x>
- Uto M, Okano M (2020) Robust neural automated essay scoring using item response theory. In: *Proceedings of the international conference on artificial intelligence in education*, pp 549–561
- Uto M, Ueno M (2016) Item response theory for peer assessment. *IEEE Trans Learn Technol* 9(2):157–170
- Uto M, Ueno M (2018) Empirical comparison of item response theory models with rater's parameters. *Heliyon* 4(5):1–32
- Uto M, Ueno M (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika* 47(2):469–496
- Uto M, Louvigné S, Kato Y, Ishii T, Miyazawa Y (2017) Diverse reports recommendation system based on latent Dirichlet allocation. *Behaviormetrika* 44(2):425–444
- Uto M, Duc Thien N, Ueno M (2020) Group optimization to maximize peer assessment accuracy using item response theory and integer programming. *IEEE Trans Learn Technol* 13(1):91–106
- van der Linden WJ (2016) *Handbook of item response theory, volume one: models*. CRC Press, Boca Raton
- van Lier HG, Siemons L, van der Laar MA, Glas CA (2018) Estimating optimal weights for compound scores: a multidimensional IRT approach. *Multivar Behav Res* 53(6):914–924
- Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput* 27(5):1413–1432
- Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res* 3571–3594. <https://doi.org/10.5555/1756006.1953045>
- Watanabe S (2013) A widely applicable Bayesian information criterion. *J Mach Learn Res* 14(1):867–897
- Wilson M, Hoskens M (2001) The rater bundle model. *J Educ Behav Stat* 26(3):283–306
- Wind SA, Jones E (2019) The effects of incomplete rating designs in combination with rater effects. *J Educ Meas* 56(1):76–100

- Yao L, Schwarz RD (2006) A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Appl Psychol Meas* 30(6):469–492
- Zhang A, Xie X, You S, Huang X (2011) Item response model parameter estimation based on Bayesian joint likelihood Langevin MCMC method with open software. *Int J Adv Comput Technol* 3(6):48–56
- Zlatkin-Troitschanskaia O, Shavelson RJ, Schmidt S, Beck K (2019) On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br J Educ Psychol* 89(3):468–484

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.