Measurement 151 (2020) 107062

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Polytomous item explanatory IRT models with random item effects: Concepts and an application

Jinho Kim *, Mark Wilson #

Graduate School of Education, University of California at Berkeley, Berkeley, USA

ARTICLE INFO

Article history:
Received 22 September 2018
Received in revised form 9 September 2019
Accepted 14 September 2019
Available online 9 October 2019

Keywords:
Item explanatory model
Polytomous data
Many-Facet Rasch Model
Linear Partial Credit Model
Linear Logistic Test Model with item error
Random item effects model
Crossed random effects

ABSTRACT

This paper proposes three polytomous item explanatory models with random item errors in Item Response Theory (IRT), by extending the Linear Logistic Test Model with item error (LLTM + ε) approach to polytomous data. The proposed models, also regarded as polytomous random item effects models, can take the uncertainty in explanation and/or the random nature of item parameters into account for polytomous items. To develop the models, the concepts and types of polytomous random item effects are investigated and then added into the existing polytomous item explanatory models. For estimation of the proposed models with crossed random effects for polytomous data, a Bayesian inference method is adopted for data analysis. An empirical example demonstrates practical implications and applications of the proposed models to the Verbal Aggression data. The empirical findings show that the proposed models with random item errors perform better than the existing models without random item errors in terms of the goodness-of-fit and reconstructing the step difficulties and also demonstrate methodological and practical differences of the proposed models in interpreting the item property effects in each of the item location explanatory Many-Facet Rasch Model and the step difficulty explanatory Linear Partial Credit Model approaches.

1. Introduction

The primary role of educational measurement and assessment is to provide necessary information about ways to facilitate teachers' instruction and students' learning [6]. This in turn emphasizes the capability and quality of informative measurement and assessment. Explanatory measurement [15] provides various explanatory inferences from the assessments so that it can strengthen the kind of feedback that could be given to both teachers and students as well as to test developers and educational researchers. In item response theory (IRT), explanatory item response models (EIRM; [15]) aim to explain the person and/or item side of the assessment data in order to enrich inferential information and enhance the feedback. Among person explanatory, item explanatory, and doubly explanatory models of the EIRM approach, this paper will focus on item explanatory models in which item properties are incorporated to explain and predict the item effects. In measurement and assessment practices, item explanatory models have various methodological advantages in extracting essential and meaningful elementary components, testing constructs hypothesized in item design and item generation, and predicting item difficulties of newly developing items as well as in measuring the effect of various testing conditions such as item presentation position, item exposure time, and testing occasions [13,19,40,58]. This item explanatory approach is also useful to examine the effect of item properties such as item design variables, item response format, content-specific learning, task characteristics, and cognitive operations in various assessment contexts [31,40]. Thus, item explanatory models can serve to provide useful and practical information for enhancing item design, item generation, and test development in educational measurement and assessment.

A typical approach to item explanatory modeling is the Linear Logistic Test Model (LLTM; [21]), which decomposes the difficulties of specific items into linear combinations of elementary components related to item properties or features [18]. The original LLTM approach has an underlying assumption that predictors of the observed item properties can perfectly account for item difficulties. However, "perfect" explanation is hardly possible because substantive theories behind the measurement model may not be flawless and/or the item difficulty parameter may be a random variable by nature [16,35]. Considering the uncertainty in

* Corresponding author at: Graduate School of Education, University of California, Berkeley, 4102 Berkeley Way Building, Berkeley, CA 94720, USA.
E-mail addresses: potatopaul@berkeley.edu (J. Kim).
Graduate School of Education, University of California, Berkeley, 4415 Berkeley Way Building, Berkeley, CA 94720, USA. markw@berkeley.edu (M. Wilson).

https://doi.org/10.1016/j.measurement.2019.107062
0263-2241/

explanation and/or the random nature of item parameters, as in an ordinary regression model, it is reasonable to add a random error or residual term into the item regression component of item explanatory models. This approach is the Linear Logistic Test Model with item error (LLTM + ε; [50,35]), which enhances prediction of the item difficulty parameters of a Rasch model by allowing for residual variation [16,31]. That is, an item error term in the LLTM + ε can compensate for the discrepancy between the freely estimated item difficulty in the Rasch model and the item difficulty calculated by the estimated item property effects in the LLTM.

Although item explanatory models have the methodological advantages and diverse potential uses, most of their applications have investigated dichotomous data (e.g., [16,17,31,34,40,58]). In particular, the LLTM and the LLTM + ε are existing and widely used item explanatory approaches to dichotomous items. In a wide range of educational, psychological, and sociological measurement and assessment contexts, however, it is common to have ordered-category responses which are regarded as polytomous data [15]. For example, an educational assessment is often developed under a learning progression framework, which generates the sophisticatedly different and ordered levels of student achievement. In measurement practices, partial credit items and rating scale items are frequently used item types, which are typically scored as ordered-category responses. Therefore, extensions and applications of item explanatory models to polytomous data, referred to as polytomous item explanatory models, need to be further investigated.

To develop polytomous item explanatory models, it is reasonable and appealing to extend the LLTM + ε approach to polytomous data with consideration for the uncertainty in explanation and/or the random nature of item parameters. For polytomous extensions of the LLTM + ε, two steps of item explanatory modeling are required: (1) item explanatory extensions of polytomous item response models using the LLTM approach—polytomous item explanatory models, (2) conceptualizing polytomous random item effects and adding them as item error terms to the item regression component of polytomous item explanatory models using the LLTM + ε approach—polytomous item explanatory models with random item effects.

For the first modeling step, polytomous item parameters in polytomous item response models need to be reparameterized by incorporating item properties, which is complicated. However, a few studies have investigated polytomous extensions of the LLTM approach. Glas and Verhelst [29] imposed linear restrictions on the item parameters of a polytomous item response model, but their reparameterization necessarily requires a complicated translation to interpret the estimated item parameters. Linacre [43] decomposed the polytomous item parameters into a linear combination of the effects of facets but the facet effects models haven't been used for item explanatory modeling and continuous item predictors cannot be incorporated in the models. Fischer and Parzer [22] and Fischer and Ponocny [23] extended the LLTM approach to polytomous item response models by using a normalization constant and basic parameters for item parameterization, however, these item parameters are difficult to interpret, and the models are complicated to incorporate item properties. Meanwhile, Kim [39] investigated item explanatory extensions of polytomous item response models under a general statistical modeling framework recently, building on the previous studies. Two polytomous item explanatory models were proposed using different item explanatory approaches to polytomous data, and the two models showed methodological and practical differences in incorporating item properties and interpreting their effects. In this paper, these two polytomous item explanatory models are employed for the first step of incremental extensions.

The second step is our main concern of item explanatory modeling for polytomous data. This modeling step to add item error terms seems straightforward, however, it is rather difficult so that polytomous extensions of the LLTM + ε approach and their applications have hardly been investigated. The difficulty originates from mainly two issues regarding polytomous random item effects: a conceptual issue and an application issue. In IRT, the LLTM + ε is a type of random item effects model in which items are treated as random and item difficulties are regarded as random effects and hence the item difficulty parameter is a random variable [16]. The LLTM assumes that the item difficulties are perfectly predicted by the fixed item property effects, whereas the LLTM + ε relaxes this assumption by allowing for random variation across items. Random item effects models are a rather new area in educational and psychological measurement research [16]. In particular, the concepts of random item effects involve a random error interpretation for the uncertainty in explanation and a random sampling interpretation for the random nature of item parameters [35]. Since these two interpretations are two sides of the same coin, polytomous random item effects and their distributional assumptions should be investigated to add item error terms into the polytomous item explanatory models. However, random item effects for polytomous items have been barely examined and/or conceptualized. A few studies have discussed them (e.g., [36,55,56]), but they have mainly investigated item selection techniques or item family calibration methods for polytomous items rather than underlying distributions for polytomous random item effects.

For applications of polytomous random item effects models in practice, it is necessary to figure out different types of polytomous item explanatory models with random item errors and to select a model between them. An overarching framework that summarizes the polytomous item explanatory models with random item errors is helpful to facilitate the understanding and applications of them, however, it has not yet been investigated. Furthermore, treating both items and persons as random makes for crossed random effects [16,35]. Estimation of random item effects models with the crossed random effects is demanding due to the complexity and difficulty in numerical integration [9,69]. Such demanding estimation becomes more difficult for polytomous data practically due to a lack of statistical software which can estimate polytomous random item effects models.

This research aims to develop and apply polytomous item explanatory models with random item errors by extending the LLTM + ε approach to polytomous data, considering the uncertainty in explanation and/or the random nature of item parameters. To specify the models, the two modeling steps of incremental extensions will be discussed in the following sections. For the first step, we will review the existing models for polytomous item explanatory extensions. Building up on polytomous item response models, the two polytomous item explanatory models that Kim [39] suggested are described. For the second step, we will examine the concepts and types of polytomous random item effects in terms of a random sampling interpretation, which enables to figure out the underlying distributions of random item errors on the polytomous item parameters. Then, we will add those random item errors to the polytomous item explanatory models in terms of a random error interpretation. Next, in addition to summarizing an overarching framework of the polytomous item explanatory models with random item errors, we will also discuss estimation methods for those models. Lastly, we will demonstrate an empirical application of the proposed models to the Verbal Aggression data to show their practical implications, interpretations, and methodological advantages.

2. Item explanatory extensions of polytomous item response models (Existing models)

2.1. Polytomous item response models

This section reviews existing polytomous item response models for the first modeling step of polytomous item explanatory extensions. Given a context of ordered-category responses regarded as polytomous data, adjacent-categories logits are employed for polytomous item response models and their item explanatory extensions. Since most of the ordered-category responses in educational assessment or cognitive development contexts are subjectively assigned scores between categories, adjacent-categories logits are suitable for modeling those scores in ordinal regression models [3,37]. Moreover, adjacent-categories logits are useful when contrasting probabilities of the responses in pairs of adjacent categories [32,49], regarded as the local comparison of ordered categories [48].

Rasch family models for polytomous data, which are based on the adjacent-categories logits, can make person and item parameters separable and hence have sufficient statistics as well as make possible specifically objective comparisons of persons and items [46,48]. These are fundamental measurement properties of the Rasch family models [47,61]. In particular, the Partial Credit Model (PCM; [46]) is a straightforward application of the Rasch model [59] originally for dichotomous responses to pairs of adjacent categories in a sequence in ordered-category responses [48]. Hence, we will use the PCM as a basic polytomous item response model to investigate random item effects in polytomous items and add random item errors to polytomous item explanatory models.

To facilitate interpretation of polytomous item parameters, explanatory convenience, and simplicity of model specification, the PCM is expressed in terms of the local comparison between adjacent categories. A comparison of the response probabilities that person p of ability θ_p scores from $m-1$ to m on item $i(i=1,2,\cdots,\mathrm{I})$, which is regarded as the m-th adjacent-categories logit [68], can be written as:

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)}=\theta_p-\delta_{im},\quad m=1,\cdots,M_i \tag{1}$$

where $\theta_p\sim N(0,\ \sigma_\theta^2)$ and $\delta_{i0}=0$. For statistical model identification[1], the mean person ability is constrained to zero. This model constraint will be consistently used for all IRT models in this paper. From this local comparison perspective of the PCM, the item parameter δ_{im} is interpreted as a "step" difficulty when shifting a category score from $m-1$ to m on item i, which is regarded as the m-th step of switching over to the next response category. Thus, a step difficulty parameter δ_{im} indicates the relative difficulty for scoring m rather than $m-1$ on item i [46,47]. Note that in this formula, the person abilities θ_p are treated as random and the step difficulties δ_{im} are treated as fixed.

In particular, the step difficulty parameter δ_{im} can be split into two item parameters, the item location (a.k.a. overall item difficulty) parameter β_i and the step deviation parameter τ_{im}, as follows:

$$\delta_{im}=\beta_i+\tau_{im}, \tag{2}$$

so that

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)}=\theta_p-\beta_i-\tau_{im},\quad m=1,\cdots,M_i \tag{3}$$

where $\theta_p\sim N(0,\ \sigma_\theta^2)$, $\tau_{i0}=0$, and $\sum_{m=1}^{M_i}\tau_{im}=0$ so that $\frac{1}{M_i}\sum_{m=1}^{M_i}\delta_{im}=\beta_i$. Due to the model constraint that $\frac{1}{M_i}\sum_{m=1}^{M_i}\delta_{im}=\beta_i$ (a mean of step difficulties for an item), an item location parameter β_i is interpreted as the overall item difficulty for polytomous item i, and a step deviation parameter τ_{im} is interpreted as a deviation of the step difficulty from the overall item difficulty for the m-th step within item i ($\tau_{im}=\delta_{im}-\beta_i$). M_i-1 step deviation parameters are estimated for each item because the sum of step deviations is constrained to zero for each item ($\sum_{m=1}^{M_i}\tau_{im}=0$). This twofold item parameterization is used in the IRT software ConQuest [1], which is essential to identify overall item difficulties, scale thresholds, and step difficulties in polytomous items. It is also useful in the context of incorporating item properties and adding an error term in order to explain the overall item difficulties.

In addition, the Rating Scale Model (RSM; [4]) is a special case of the PCM [46], in which the relative difficulties of the steps between categories do not vary across all items ($\tau_{im}=\tau_m$) and the number of steps are the same for all items ($M_i=M$). In the RSM, the item location parameter β_i and the common scale threshold parameter τ_m can be formulated as:

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)}=\theta_p-\beta_i-\tau_m,\quad m=1,\cdots,M \tag{4}$$

where $\theta_p\sim N(0,\ \sigma_\theta^2)$, $\tau_0=0$, and $\sum_{m=1}^{M}\tau_m=0$. Note that the RSM imposes more restrictions on the step deviation parameter of the PCM. In sum, the PCM is the most general Rasch family model for polytomous (ordinal) data [61].

2.2. Polytomous item explanatory models

We will employ the existing models for polytomous item explanatory extensions. The two polytomous item explanatory models that Kim [39] suggested are reviewed in this section. The two models are item explanatory extensions of the PCM, which are recently investigated building on the previous studies, and their parameters of item property effects are straightforward to interpret. For the purpose of developing polytomous item explanatory models with random item errors in the second modeling step, it is necessary to figure out the two polytomous item explanatory approaches before adding random item errors. As discussed in the introduction, however, reparameterizing polytomous item parameters by incorporating item properties into the PCM is complicated (also see [39]). To avoid the complication of polytomous item parameterization, it is helpful to clarify the target difficulty parameters in the PCM which are explained by item properties: (a) the item location (overall item difficulty) parameters β_i can be explained by item properties, and (b) the step difficulty parameters δ_{im} can be explained by item properties.

2.2.1. Item location explanatory Many-Facet Rasch Model (MFRM)

In the first case, by using the twofold item parameterization for the PCM in Eq. (3), we can impose linear restrictions on the item location parameters as was done in the LLTM, and also estimate the step deviation parameters for each item. The restricted item location parameters β_i' are decomposed into weighted sums of item property effect parameters γ_k as follows:

$$\beta_i'=\sum_{k=0}^{K}\gamma_k x_{ik},\quad k=0,\cdots,K \tag{5}$$

so that

$$ln\frac{Pr(y_{pi}=m|\theta_p)}{Pr(y_{pi}=m-1|\theta_p)}=\theta_p-\sum_{k=0}^{K}\gamma_k x_{ik}-\tau_{im} \tag{6}$$

[1] For model identification, a particular item parameter or the mean of item parameters can be constrained to zero. In this case, the mean person ability is estimated [15].

where $\theta_p \sim N(0,\ \sigma_\theta^2)$, $\tau_{i0} = 0$, and $\sum_{m=1}^{M_i}\tau_{im} = 0$. Also, where γ_0 is the item intercept representing the difficulty for items with all x_{ik} = 0 for k > 0, γ_k is the regression weight or the effect of item property k on the overall item difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, x_{ik} is the predictor value of item i on item property k, and τ_{im} is a step deviation parameter for the m-th step of item i.. Note that in general the restricted item location parameters $\beta_i^{'}$ in Eq. (5) will not be equal to the original parameters β_i in Eq. (3) because the explanation by the observed item properties are not perfect.

This item location explanatory extension of the PCM is a variation of Linacre [43]'s Many-Facet Rasch Model (MFRM) in that the item location parameters are restricted by a linear combination of the effects of item properties which are regarded as sub-facets within the item facet [75]. Building on the MFRM, a more general regression notation is used to include both categorical and continuous explanatory predictors, and the step deviation parameters are estimated for each item. This polytomous item explanatory model with a decomposition of the item location parameters will be called the "*item location explanatory* Many-Facet Rasch Model (MFRM)".

2.2.2. Step difficulty explanatory Linear Partial Credit Model (LPCM)

For the second case, we can impose linear restrictions on the step difficulty parameters in Eq. (1) by incorporating item properties. The restricted step difficulty parameters $\delta_{im}^{'}$ are decomposed into weighted sums of step specific item property effect parameters ω_{km} as follows:

$$\delta_{im}^{'} = \sum_{k=0}^{K}\omega_{km}x_{ik}, \quad k = 0, \cdots, K \tag{7}$$

so that

$$ln\frac{Pr(y_{pi} = m|\theta_p)}{Pr(y_{pi} = m-1|\theta_p)} = \theta_p - \sum_{k=0}^{K}\omega_{km}x_{ik} \tag{8}$$

where $\theta_p \sim N(0,\ \sigma_\theta^2)$, $\omega_{k0} = 0$, and also where ω_{0m} is the step intercept representing the m-th step difficulty for items with all x_{ik} = 0 for k > 0, ω_{km} is the regression weight or the effect of item property k on the m-th step difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, and x_{ik} is the predictor value of item i on item property k. Note that in general the restricted step difficulty parameters $\delta_{im}^{'}$ in Eq. (7) will not equal the original parameters δ_{im} in Eq. (1) because there is no complete explanation via the observed item properties.

This step difficulty explanatory extension of the PCM is a variation of the Linear Partial Credit Model (LPCM; [23]) in that the step difficulty parameters are restricted by a linear combination of item properties, although they used complicated and less interpretable item parameters such as a normalization constant and basic parameters. Building on the LPCM, the step specific item property effect parameters are used to interpret the effects of item properties on the step difficulties for each step. This polytomous item explanatory model with a decomposition of the step difficulty parameters will be called the "*step difficulty explanatory* Linear Partial Credit Model (LPCM)".

3. Polytomous item explanatory models with random item effects (Proposed models)

3.1. Polytomous random item effects

For the second modeling step, the concepts and types of random item effects for polytomous items should be investigated to incorporate random item errors into the two polytomous item explanatory models. The concepts of random item effects are related to the random nature of item parameters in terms of a random sampling interpretation as well as to the uncertainty in explanation in terms of a random error interpretation [35]. Since random item effects parameters are the same random variables with the same distributional assumptions in both interpretations, these two ways of interpreting random item effects are two sides of the same coin. In this section, we will first examine random item effects that account for the random nature of polytomous item parameters in the first interpretation and then will address how to incorporate them in the second interpretation.

The notion of random item effects becomes clearer in the context of item banks, item generation (or item cloning), and automatic item generation (AIG) as well as generalizability in item response modeling (GIRM). In practice, there are some preexisting item populations. For instance, with the advent of computer adaptive testing (CAT), item banks or item pools are created using IRT models and they are used as the item population for random item selection [16,28,30]. In CAT, items are randomly drawn from an item bank so that a homogenous set of the items can be constructed. Also, item generation is seen as formally equal to drawing from a theoretical population such as in a theory of domain-referenced testing [16,33]. The concept of "domain", also referred to as "universe", means the knowledge and skills required for the mastery of a specific content area in educational assessment [35]. In AIG, items (or item clones) within an item family are generated automatically based on the item generating models and are administered on a computer-based test [34,36,41]. The concept of "item family", based on the principal of item generation, represents a set of items with sufficient commonalities as well as sufficient differentiation from other sets of items [16,28,62]. Such an item universe or item family is regarded as the item population from which test items are randomly sampled or automatically generated. This implies the items are seen as random. Moreover, the random nature of the items concerns the generalizability potential of a random item effects model [16]. As in the GIRM approach [8,12], generalization to the item universe can be a matter of concern even when the items are not randomly sampled in reality from a preexisting item population. Thus, random item effects models are promising and useful for the calibration of item parameters from the underlying item population in various measurement situations.

Despite the practical implications and the potential usefulness in educational measurement practices, random item effects models are a rather new approach in IRT [16]. In fact, little is known about random item effects for polytomous items. Most relevant studies that have investigated random item effects have used dichotomous data (e.g., [9,16,17,28,31,34]). Although a few studies have discussed polytomous random item effects (e.g., [36,55,56]), they have mainly investigated item selection techniques or item family calibration methods for polytomous items in the context of CAT and AIG rather than underlying distributional assumptions for polytomous random item effects. It has been rare to investigate and/or conceptualize random item effects in polytomous item response models.

In the random item effects model setting, items are treated as random. The randomness implies that there could, in principle, be many items. In the item response process for polytomous (ordinal) items, the scale structure is fixed by a scale type and outcome space corresponding to constructs for item design such as a learning progression; thus, the number of response categories is fixed and hence that of category steps is fixed. In order to come up with polytomous random item effects in terms of the random sampling interpretation, this paper focuses on the polytomous item parameters of the PCM. To allow for random item effects in the PCM, it is helpful to examine the conceptualization of random effects on the

item side of the polytomous responses by considering different types of item parameters. To our knowledge, there are three types of polytomous random item effects: (a) overall item random effects—random variations in overall item difficulties across items, (b) multivariate item-step random effects—random variations in step difficulties across items and category steps and they are correlated between category steps, and (c) univariate item-step random effects—random variations in step difficulties across items and category steps and they are uncorrelated between category steps.

For the first type of polytomous random item effects, by using the twofold item parameterization for the PCM in Eq. (2), random effects can be imposed on the overall item difficulties for individual items so that the item location parameters β_i are a random variable, expressed as:

$$\delta_{im} = \beta_i + \tau_{im} \text{ and } \beta_i \sim N\left(\mu_\beta,\ \sigma_\beta^2\right) \tag{9}$$

where $\theta_p \sim N(0,\ \sigma_\theta^2)$, $\tau_{i0} = 0$ and $\sum_{m=1}^{M_i} \tau_{im} = 0$. The overall item random effects, denoted by β_i, are assumed to follow a normal distribution with a mean of μ_β and a variance of σ_β^2. Due to the model constraint that $\theta_p \sim N(0,\ \sigma_\theta^2)$, μ_β represents a grand mean overall item difficulty across items over persons. In addition, the sum of step deviations is constrained to zero for each item ($\sum_{m=1}^{M_i} \tau_{im} = 0$) to identify the overall difficulty of an item, and the step deviation parameter τ_{im} is not random. In this approach, the explanation and prediction of the *overall* item difficulties are a matter of interest in interpretation of random item effects.

In the second type of polytomous random item effects, the explanation and prediction of the step difficulties are important. Given the same number of response categories and category steps from the same scale structure in the same item population, the step difficulties can vary within an item and between items as the items vary. In the PCM, the step difficulties of an item do not rely on the particulars of the other items in the sample from the item population and there are no order restrictions between them within an item [14,46,48]. However, within an item, the step difficulties are not interpreted independently of each other [47,51]. That is, category scores have some deterministic dependence on the relative difficulties between category steps in the same item despite the assumption of local independence between items given the person ability [46,52]. This implies a distribution assumption for the ordinal item population that the step difficulties in each step can randomly vary across items but they are correlated within an item.

To accommodate these concerns, by using the original PCM's item parameterization in Eq. (1), correlated item-step random effects between category steps can be imposed on the step difficulties step-specifically for individual items so that the step difficulty parameter vector $\boldsymbol{\delta_i}$ are a set of correlated random variables, as follows:

$$\boldsymbol{\delta_i}\ (\delta_{i1},\ \delta_{i2},\ \ldots,\ \delta_{im})' \sim MVN_m\left(\boldsymbol{\mu_\delta},\ \boldsymbol{\Sigma}\right) \tag{10}$$

where $\boldsymbol{\mu_\delta} = \begin{bmatrix} \mu_{\delta 1} \\ \mu_{\delta 2} \\ \vdots \\ \mu_{\delta m} \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\delta 1}^2 & \sigma_{\delta 1 \delta 2} & \cdots & \sigma_{\delta 1 \delta m} \\ \sigma_{\delta 2 \delta 1} & \sigma_{\delta 2}^2 & \cdots & \sigma_{\delta 2 \delta m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\delta m \delta 1} & \sigma_{\delta m \delta 2} & \cdots & \sigma_{\delta m}^2 \end{bmatrix}$, $\theta_p \sim N(0,\ \sigma_\theta^2)$, and $\delta_{i0} = 0$. The vector of the multivariate item-step random effects $\boldsymbol{\delta_i}\ (\delta_{i1}, \delta_{i2}, \cdots, \delta_{im})'$ is assumed to follow a multivariate normal (MVN) distribution with a mean vector $\boldsymbol{\mu_\delta}\ (\mu_{\delta 1}, \mu_{\delta 2}, \cdots, \mu_{\delta m})'$ and a variance–covariance matrix $\boldsymbol{\Sigma}(m \times m)$. Due to the model constraint that $\theta_p \sim N(0,\ \sigma_\theta^2)$, $\mu_{\delta m}$ represents a grand mean step difficulty for the m-th step across items over persons.

The third type of polytomous random item effects are univariate item-step random effects that the step difficulties can randomly vary across items and category steps. By using the original PCM's item parameterization in Eq. (1), uncorrelated item-step random effects between category steps can be imposed on the step difficulties for individual items so that the step difficulty parameters δ_{im} are a random variable, written as:

$$\delta_{im} \sim N(\mu_\delta,\ \sigma_\delta^2) \tag{11}$$

where $\theta_p \sim N(0,\ \sigma_\theta^2)$ and $\delta_{i0} = 0$. The univariate item-step random effects δ_{im} is assumed to follow a normal distribution with a mean of μ_δ and a variance of σ_δ^2. Due to the model constraint that $\theta_p \sim N(0,\ \sigma_\theta^2)$, μ_δ represents a grand mean step difficulty for all category steps across items over persons.

In fact, this univariate item-step random effects model is a special case of the multivariate item-step random effects model in Eq. (10). The two models are identical with additional assumptions for the latter that the mean and variance of the item-step random effects are the same across every step ($\mu_{\delta 1} = \cdots = \mu_{\delta m} = \mu_\delta$ in $\boldsymbol{\mu_\delta}$; $\sigma_{\delta 1}^2 = \cdots = \sigma_{\delta m}^2 = \sigma_\delta^2$ in $\boldsymbol{\Sigma}$) and the item-step random effects are independent between category steps (zero correlation or zero covariance; $\sigma_{\delta 1 \delta 2} = \cdots = \sigma_{\delta m \delta m-1} = 0$ in $\boldsymbol{\Sigma}$). The univariate item-step random effects seem to be peculiar due to assumptions that do not reflect the multi-categorical scale structure in the ordinal items. Nevertheless, this type of polytomous random item effects has been commonly used for a prior distribution of the step difficulty parameters in Bayesian estimation methods (e.g., [2,14,24,38]). This makes sense in that a weakly or non-informative prior for model parameters need not be strictly true in Bayesian inference [24,65].

The three types of polytomous random item effects for the PCM are summarized with each distribution assumption in Table 1. Since the two interpretations of random item effects are closely related, a distribution assumption for each polytomous random item effects model will be retained for corresponding random item errors in the following section. In brief, for the overall item random effects, the item location parameter is a random variable that follows a univariate normal distribution and the explanation and prediction of the overall item difficulties are a matter of interest regardless of the ordinal scale structure. For the multivariate item-step random effects, the set of step difficulty parameters is

Table 1
Three types of polytomous random item effects for the PCM.

Item Parameterization	Random Item Parameter	Polytomous Random Item Effects	Distribution Assumption
Two fold $\delta_{im} = \beta_i + \tau_{im}$	Item Location Parameter β_i	Overall Item Random Effects	$\beta_i \sim N\left(\mu_\beta, \sigma_\beta^2\right)$ and $\sum_{m=1}^{M_i} \tau_{im} = 0$
One fold δ_{im}	Step Difficulty Parameter δ_{im}	Multivariate Item-Step Random Effects	$\boldsymbol{\delta_i}(\delta_{i1},\ \ldots,\ \delta_{im})' \sim MVN_m(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$
One fold δ_{im}	Step Difficulty Parameter δ_{im}	Univariate Item-Step Random Effects	$\delta_{im} \sim N(\mu_\delta, \sigma_\delta^2)$

Note. The overall item random effectsβ_i follow a normal distribution with a mean of μ_β and a variance of σ_β^2. The vector of the multivariate item-step random effects $\boldsymbol{\delta_i}$ $(\delta_{i1}, \delta_{i2}, \cdots, \delta_{im})'$ follows a multivariate normal distribution with a mean vector $\boldsymbol{\mu_\delta}$ $(\mu_{\delta 1}, \mu_{\delta 2}, \cdots, \mu_{\delta m})'$ and a variance-covariance matrix $\boldsymbol{\Sigma}(m \times m)$. The univariate item-step random effects δ_{im} follow a normal distribution with a mean of μ_δ and a variance of σ_δ^2.

a random vector that follows a multivariate normal distribution and the explanation and prediction of the correlated step difficulties between category steps are a matter of interest with consideration for the ordinal scale structure. For the univariate item-step random effects, the step difficulty parameter is a random variable that follows a univariate normal distribution and the explanation and prediction of the uncorrelated step difficulties between category steps are a matter of interest without consideration for the ordinal scale structure.

In addition, the multivariate item-step random effects model in Eq. (10) hints at a random item effects version of the RSM. As in the overall item random effects model in Eq. (9), the item location parameters β_i can be a random variable which follows a normal distribution with a mean of μ_β and a variance of σ^2_β. Based on the RSM assumption, the relative difficulties between category steps are the same for all items so that the common scale threshold parameter τ_m in Eq. (4) concerns only the rating scale structure. By summing up μ_β and τ_m, one can identify a similar mean vector with the multivariate item-step random effects model. A further investigation for this topic is needed but we will leave it for a future study in order to focus on extensions of the PCM in this paper.

3.2. Polytomous item explanatory models with random item errors

This section addresses how to develop polytomous item explanatory models with random item errors for the second step of polytomous extensions of the LLTM + ε approach. We will incorporate the three types of polytomous random item effects into the two polytomous item explanatory models in a random error interpretation of random item effects. The polytomous random item effects have an underlying distribution which can be used for each corresponding type of polytomous random item errors. As in the LLTM + ε approach, an error term can enhance prediction of the polytomous item difficulties, the overall item difficulties and the step difficulties, by allowing for residual variation.

Note that the two polytomous item explanatory models are methodologically and functionally different item explanatory approaches to polytomous items in terms of the target item parameters of the polytomous item difficulties which are explained by item properties and the types of item property effects [39]. Moreover, the three types of polytomous random item errors are conceptually different in terms of the random item parameters, the underlying distributions, and the consideration for the ordinal scale structure. To specify polytomous item explanatory models with random item errors, it is necessary to clarify the target polytomous item difficulty parameters which are explained by item properties as well as the types of polytomous random item errors which are unexplained residuals. First, in the item location explanatory MFRM approach, the item location parameters β_i can be explained by item properties and then (a) overall item random error can be added for unexplained residual variation in the overall item difficulties regardless of the ordinal scale structure. Next, in the step difficulty explanatory LPCM approach, the step difficulty parameters δ_{im} can be explained by item properties and then (b) multivariate item-step random error or (c) univariate item-step random error can be added for unexplained residual variation in the step difficulties, depending on the consideration for the ordinal scale structure.

3.2.1. Item location explanatory Many-Facet Rasch Model with Overall Item Random Error (MFRM + OIE)

The first model is referred to as the *item location explanatory* Many-Facet Rasch Model with Overall Item Random Error (MFRM + OIE). By using the twofold item parameterization for the PCM, an overall item random error term ϵ_i is added to the MFRM-based polytomous item explanatory model in Eqs. (5) and (6), formulated as:

$$\delta_{im} = \sum_{k=0}^{K} \gamma_k x_{ik} + \epsilon_i + \tau_{im} \tag{12}$$

where $\epsilon_i \sim N(0, \sigma^2_\epsilon)$, $\theta_p \sim N(0,\ \sigma^2_\theta)$, $\tau_{i0} = 0$, and $\sum_{m=1}^{M_i} \tau_{im} = 0$. Also, where ϵ_i is an error or residual term for the overall item difficulties with a normal distribution of a zero mean and a variance of σ^2_ϵ, γ_0 is the item intercept representing the difficulty for items with all x_{ik} = 0 for k > 0, γ_k is the regression weight or the effect of item property k on the overall item difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, x_{ik} is the value of item i on item property k, and τ_{im} is a step deviation parameter for the m-th step of item i. Note that this model with no item property predictors ($K = 0$) is formally equivalent to the overall item random effects model in Eq. (9), that is, $\gamma_0 = \mu_\beta$.

3.2.2. Step difficulty explanatory Linear Partial Credit Model with Multivariate Item-Step Random Error (LPCM + MISE)

The second model is referred to as the *step difficulty explanatory* Linear Partial Credit Model with Multivariate Item-Step Random Error (LPCM + MISE). By using the original PCM's item parameterization, a vector of the multivariate item-step random error term $\boldsymbol{\xi_i}(\xi_{i1}, \xi_{i2}, \cdots, \xi_{im})'$ is added to the LPCM-based polytomous item explanatory model in Eqs. (7) and (8), formulated as:

$$\delta_{im} = \sum_{k=0}^{K} \omega_{km} x_{ik} + \xi_{im} \tag{13}$$

where $\boldsymbol{\xi_i} = \begin{bmatrix} \xi_{i1} \\ \xi_{i2} \\ \vdots \\ \xi_{im} \end{bmatrix} \sim MVN_m(\mathbf{0},\ \boldsymbol{\Sigma}),\ \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$,

$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{\xi 1} & \sigma_{\xi 1 \xi 2} & \cdots & \sigma_{\xi 1 \xi m} \\ \sigma_{\xi 2 \xi 1} & \sigma^2_{\xi 2} & \cdots & \sigma_{\xi 2 \xi m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\xi m \xi 1} & \sigma_{\xi m \xi 2} & \cdots & \sigma^2_{\xi m} \end{bmatrix}$, $\theta_p \sim N(0,\ \sigma^2_\theta)$, and $\omega_{k0} = 0$. Also, where $\boldsymbol{\xi_i}$ is an error or residual term vector for the step difficulties following a multivariate normal (MVN) distribution with a zero mean vector $\mathbf{0}$ and a variance–covariance matrix $\boldsymbol{\Sigma}(m \times m)$, ω_{0m} is the step intercept representing the m-th step difficulty for items with all x_{ik} = 0 for k > 0, ω_{km} is the regression weight or the effect of item property k on the m-th step difficulties, x_{io} is the constant item predictor in which a value of 1 for all items, and x_{ik} is the value of item i on item property k. Note that this model with no item property predictors ($K = 0$) is formally equivalent to the multivariate item-step random effects model in Eq. (10), that is, $\omega_{0m} = \mu_{\delta m}$ for every m-th step.

3.2.3. Step difficulty explanatory Linear Partial Credit Model with Univariate Item-Step Random Error (LPCM + UISE)

The third model is referred to as the *step difficulty explanatory* Linear Partial Credit Model with Univariate Item-Step Random Error (LPCM + UISE). By using the original PCM's item parameterization, a univariate item-step random error term ε_{im} is added to the LPCM-based polytomous item explanatory model in Eqs. (7) and (8), formulated as:

$$\delta_{im} = \sum_{k=0}^{K} \omega_{km} x_{ik} + \varepsilon_{im} \tag{14}$$

where $\varepsilon_{im} \sim N(0, \sigma_{\varepsilon}^2)$, $\theta_p \sim N(0, \sigma_{\theta}^2)$, and $\omega_{k0} = 0$. Also, where ε_{im} is an error or residual term for the step difficulties with a normal distribution of a zero mean and a variance of σ_{ε}^2, and other parameters are defined as before in the second model. Note that this model with no item property predictors ($K = 0$) is not equivalent to the univariate item-step random effects model in Eq. (11), that is, $\omega_{0m} \neq \mu_{\delta}$. To be identical, values of the step intercepts should be collapsed across steps into the one same value ($\omega_{01} = \cdots = \omega_{0m} = \omega_0 = \mu_{\delta}$). This is because the univariate item-step random effects do not adequately reflect the multi-categorical scale structure in the ordinal items. Nonetheless, the univariate item-step random effects can be used to allow for residual variation in the step difficulties. The multivariate item-step random effects are more general than the univariate item-step random effects which have additional assumptions of the same mean and variance for every step and the zero correlation between steps. In this sense, the LPCM + MISE is a more general LPCM-based polytomous item explanatory model than the LPCM + UISE.

Table 2 presents an overarching framework summarizing the polytomous item explanatory models with random item errors that we propose with model specifications in this paper. In sum, the three proposed models are functionally and conceptually different polytomous item explanatory models in terms of (1) the target polytomous item difficulty parameters which are explained by item properties, (2) the types of item property effects, and (3) the types of polytomous random item errors which are unexplained residuals. In applications of the proposed models to polytomous data, these methodological differences between the models may lead to practical differences such as different interpretations of the estimated item property effects and/or different considerations for the ordinal scale structure. In order to select a specific model, considering the uncertainty in explanation and/or the random nature of item parameters, it is necessary and helpful to figure out the methodological and practical differences of the proposed models as well as different measurement contexts. For example, if one may want to explain the overall item difficulties by item properties regardless of the ordinal scale structure, the MFRM + OIE would be the best model. If one may want to predict the step difficulties for each step of new items using the item property effects, considering the ordinal scale structure in unexplained residuals, the LPCM + MISE would be the most useful model. If one may want to see the effects of item properties on the step difficulties for each step without consideration for the ordinal scale structure in unexplained residuals, the LPCM + UISE would be the most suitable model.

3.3. Estimation of polytomous random item effects models

The proposed models, regarded as polytomous random item effects models, essentially have crossed random effects. Given the fact that item responses are nested in persons and also in items but these two sides of the response data are not nested, allowing for random effects on both item and person sides makes them crossed [69]. However, estimation of crossed random effects in IRT is demanding due to the complexity and difficulty in numerical integration [9,15]. Even worse, such demanding estimation becomes more practically difficult for the polytomous random item effects models which use adjacent-categories logits. Compared to a logit link, an adjacent-categories logit link is not common in general statistical software implementing maximum likelihood-based estimation methods. In practice, software is rare that can estimate crossed random effects using the adjacent-categories logit link for polytomous data. For example, the *lmer* function in the *lme4* R package [17] and the *xtmelogit/meglm* commands in Stata [67], which use the Laplace approximation [5], can fit logistic crossed random effects models to dichotomous data but cannot fit adjacent-categories logit-based crossed random effects models to polytomous data. Moreover, most of the IRT software such as PARSCALE [54] and BILOG-MG [74] cannot estimate crossed random effects IRT models.

To estimate the polytomous random item effects models, Bayesian inference is the most practical and feasible estimation method, which uses Markov chain Monte Carlo (MCMC) algorithm [24]. This simulation-based method is a straightforward approach to estimation of the crossed random effects because it considers all effects as random parameters [10]. It is also flexible and useful to estimate complicated models tailored to specific research questions [11,36]. Although Bayesian inference is not based on maximum likelihood estimation, it can yield very similar results as other comparable methods such as the alternating imputation posterior (AIP) and the Laplace approximation [9,11,16,35]. In fact, Bayesian inference method is getting to be more commonly used (e.g., [2,7,14,24,38,41,45]).

4. Empirical application to the Verbal Aggression data

4.1. Data

For an empirical example, we used the Verbal Aggression data set [70], which is publicly available at the BEAR center website page (see [15]; the data set can be downloaded from http://bearcenter.berkeley.edu/EIRM/). The data were collected from the first-year psychology students of a Belgian university.

Table 2
Three proposed polytomous item explanatory models with random item errors.

Item Parameterization	Target Item Difficulty Parameter	Relevant Explanatory Approach	Polytomous Random Item Error	Model Specification
Two fold $\delta_{im} = \beta_i + \tau_{im}$	Item Location (Overall Item Difficulty) Parameter β_i	Many-Facet Rasch Model $\beta_i^{'} = \sum_{k=0}^{K} \gamma_k x_{ik}$	Overall Item Random Error ϵ_i	$\delta_{im} = \sum_{k=0}^{K} \gamma_k x_{ik} + \epsilon_i + \tau_{im}$, $\epsilon_i \sim N(0, \sigma_{\epsilon}^2), \sum_{m=1}^{M_i} \tau_{im} = 0$
One fold δ_{im}	Step Difficulty Parameter δ_{im}	Linear Partial Credit Model $\delta_{im}^{'} = \sum_{k=0}^{K} \omega_{km} x_{ik}$	Multivariate Item-Step Random Error ξ_{im}	$\delta_{im} = \sum_{k=0}^{K} \omega_{km} x_{ik} + \xi_{im}$, $\boldsymbol{\xi_i}(\xi_{i1}, \cdots, \xi_{im})' \sim MVN_m(\mathbf{0}, \boldsymbol{\Sigma})$
			Univariate Item-Step Random Error ε_{im}	$\delta_{im} = \sum_{k=0}^{K} \omega_{km} x_{ik} + \varepsilon_{im}$, $\varepsilon_{im} \sim N(0, \sigma_{\varepsilon}^2)$

Note. x_{ik} is the value of item i on item property k, γ_k is the regression weight or the effect of item property k on the overall item difficulties, τ_{im} is a step deviation parameter for the m-th step of item i, and ω_{km} is the regression weight or the effect of item property k on the m-th step difficulties, ϵ_i is an error term for residual variation in the overall item difficulties, $\boldsymbol{\xi_i}(\xi_{i1}, \xi_{i2}, \cdots, \xi_{im})'$ is an error term vector for multivariate item-step residual variation in the step difficulties, and ε_{im} is an error term for univariate item-step residual variation in the step difficulties.

Table 3
The Verbal Aggression items and three item design factors.

Item	Situations	Behavior Mode		Situation Type		Behavior Type		
		Want	Do	Other-to-blame	Self-to-blame	Curse	Scold	Shout
1	A bus fails to stop for me.	1	0	1	0	1	0	0
2		1	0	1	0	0	1	0
3		1	0	1	0	0	0	1
4	I miss a train because the clerk gave me faulty information.	1	0	1	0	1	0	0
5		1	0	1	0	0	1	0
6		1	0	1	0	0	0	1
7	The grocery store closes just as I am about to enter.	1	0	0	1	1	0	0
8		1	0	0	1	0	1	0
9		1	0	0	1	0	0	1
10	The operator disconnects me when I used up my last 10 cents for a call.	1	0	0	1	1	0	0
11		1	0	0	1	0	1	0
12		1	0	0	1	0	0	1
13	A bus fails to stop for me.	0	1	1	0	1	0	0
14		0	1	1	0	0	1	0
15		0	1	1	0	0	0	1
16	I miss a train because the clerk gave me faulty information.	0	1	1	0	1	0	0
17		0	1	1	0	0	1	0
18		0	1	1	0	0	0	1
19	The grocery store closes just as I am about to enter.	0	1	0	1	1	0	0
20		0	1	0	1	0	1	0
21		0	1	0	1	0	0	1
22	The operator disconnects me when I used up my last 10 cents for a call.	0	1	0	1	1	0	0
23		0	1	0	1	0	1	0
24		0	1	0	1	0	0	1

They were asked to answer the behavioral questions about verbally aggressive reactions to frustrating situations. In total, 7,584 observations from 316 persons responding to the 24 Verbal Aggression items were included in the data set. For polytomous data IRT analyses, the original three ordered-category responses (no = 0, perhaps = 1, yes = 2) were used without a dichotomization.

Influenced by three experimental design factors, the Verbal Aggression items were designed to have a stem describing a frustrating situation and a verbal aggression response part. The first factor is the Behavior Mode which has two levels of behavior modes—wanting (Want) and doing (Do). For the second factor, the Situation Type has two types of situations—situations in which someone else is to blame (Other-to-blame) and situations in which oneself is to blame (Self-to-blame). The third factor, the Behavior Type, has three kinds of verbal aggressive behaviors—cursing (Curse), scolding (Scold), and shouting (Shout).

For the verbal aggression response part, a total of six responses were made by mixing the two behavior modes and the three verbal aggressive behaviors. For the frustrating situation, two situation cases were made within each situation type: two cases (bus stop, missing a train) for the other-to-blame situation and two cases (grocery store close, disconnected call by an operator) for the self-to-blame situation. A total of 24 items were written by the item stem which comprises one of the six aggressive responses and one of the four frustrating situations. By the three factorial ($2 \times 2 \times 3$) item design with two cases within each cell, the 24 Verbal Aggression items are classified by the predictors of each design factor as in Table 3.

These item design factors are regarded as categorical item properties and their predictors are functioning as weights of the elementary components gathered into a Q matrix in the LLTM approach. To incorporate the three item properties into polytomous item explanatory models, they were dummy coded; the Want in the Behavior Mode property, the Self-to-blame in the Situation Type property, and the Shout in the Behavior Type property served a reference for each item property.

4.2. Analysis

An empirical study was conducted to examine how the proposed polytomous item explanatory models with random item errors (the MFRM + OIE, the LPCM + MISE, and the LPCM + UISE) perform in practice. To evaluate the relative performance of the proposed models in comparison with the existing models, we compared them with the polytomous item explanatory models without random item errors (the MFRM and the LPCM) as well as with the saturated model in polytomous Rasch models (the PCM). In total, the six models were fitted to the Verbal Aggression data by means of a Bayesian inference method.

To implement the MCMC, the *RStan* R package [66], the R implementation of Stan [8], was used for data analysis. Stan is a newly developed Bayesian program that implements a No-U-Turn sampler (NUTS) based on a Hamiltonian Monte Carlo (HMC) sampling algorithm. NUTS is more efficient and robust than Gibbs sampling or the Metropolis-Hastings algorithm to explore the posterior parameter space [8,45]. As model size and complexity grow, by virtue of NUTS, Stan is faster and more robust than previous Bayesian software such as the WinBUGS based on Gibbs sampling [44] for models with complex posteriors[2] [45,53]. Stan provides full Bayesian inference for model parameters treated as random variables. Bayesian inference requires a set of prior distributions of the model parameters that could reflect our beliefs or a priori knowledge about those parameters before some evidence is taken into account [24]. The priors and hyperpriors of model parameters as well as the Stan codes are described for the three proposed models in Appendix.

Four chains of 1000 iterations were used for MCMC estimation. A total of 2000 draws across the four chains were used to estimate posterior means of the parameters after discarding the first 500

[2] In the pilot simulation study we conducted, as in the findings by Monnahan, Thorson, and Branch [53], Stan outperformed WinBUGS to estimate the LPCM + MISE which is the most complex polytomous random item effects model. It appeared that fitting the model to simulated data could not converge using WinBUGS with 4 chains and 7000 iterations but it could converge using Stan with 4 chains and 1000 iterations.

burn-in (or warm-up) iterations of each chain. Convergence of the four chains was monitored by checking the estimated potential scale reduction statistic $\widehat{R}$ [27]. A value of the $\widehat{R}$ statistic near 1 usually indicates model convergence. The value of 1.1 has been recommended as a threshold and $\widehat{R} < 1.1$ is desirable [27,45,65]. For additional convergence diagnostics, we also checked the effective sample size for a chain, which is an estimate of the number of independent draws that would lead to the same expected precision (see [8,65]).

The fitted models were evaluated by comparing goodness-of-fit and also by examining agreement for the estimated and calculated step difficulties. Three Bayesian goodness-of-fit indices were computed for each model: the deviance information criterion (DIC; [63,64]), the widely applicable information criterion (WAIC; [73]), and an information criterion version of the leave-one-out cross-validation (LOOIC; [71,72]). These Bayesian measures of model fit and adequacy are widely used to assess the generalization utility or the predictive ability of the candidate models [57]. In particular, WAIC and LOOIC are theoretically justified and empirically superior to traditional model selection criteria such as the Akaike information criterion [45,57,64]. We computed DIC using the log-likelihood from the Stan output, and the *loo* R package [72] was used to compute WAIC and LOOIC. It should be noted that all these methods are conditional on the person abilities and item random effects, and it might be preferable to use methods based on marginal likelihoods [26]. We also examined agreement for the estimated and calculated step difficulties between the fitted models using correlation coefficients and graphical comparisons (see [21,34]). The directly estimated step difficulties in the PCM were compared to the calculated step difficulties in the polytomous item explanatory models.

In addition, to demonstrate methodological and practical differences of the three proposed models, the estimated item property effects were reported and interpreted separately in each of the two item explanatory approaches to polytomous items. Particularly, the effects of item properties on the overall item difficulties in the item location explanatory MFRM approach and those on the step difficulties in the step difficulty explanatory LPCM approach were interpreted. A concept of the coefficient of determination (R^2) was also used to see an explanatory power of the item properties. In this paper, R^2 was calculated as the proportion of the total item variance that is explained by the item property effects:

$$R^2 = \frac{v^2_{model}}{v^2_{model} + r^2_{error}}$$

where v^2_{model} is the variance of the weighted sums of the item property effects in the item explanatory model and r^2_{error} is the item error (or residual) variance.

4.3. Results

The five polytomous item explanatory models with and without random item errors and the PCM were fitted to the Verbal Aggression data. For convergence diagnostics, the MCMC simulations had converged in that the largest $\hat{R}$ statistic value was 1.01 and the effective sample sizes were fairly large in all fitted models.

Table 4 shows the goodness-of-fit results of the six models. It was revealed that the order of DIC, LOOIC, and WAIC across the models was consistent: the LPCM + MISE, the LPCM + UISE, the MFRM + OIE, the PCM, the MFRM, and the LPCM fit better to the Verbal Aggression data in sequence (i.e., smaller values of the three goodness-of-fit indices). This result agrees with the findings from a simulation study that the three random item effects models showed a superior goodness-of-fit to the PCM in the high information (R^2= 0.9) condition when the simulation data were generated using the LPCM + MISE [39]. This hints that random item errors in the three proposed models could fully account for residual variation in the polytomous item difficulties of the Verbal Aggression items and/or the three factorial item design worked well for item generation in both the item location explanatory MFRM and the step difficulty explanatory LPCM approaches. It is demonstrated that the three item properties (design factors) has a high explanatory value for the Verbal Aggression items, as high R^2 values were calculated below in each of the two approaches.

The results reveal that the MFRM and the LPCM fit worse than the PCM and the MFRM fit better than the LPCM. These are as expected because, in general, item explanatory models without random item errors fit worse than the saturated model and because the MFRM had the larger number of parameters than the LPCM.

We looked at practical significance by examining the agreement between the directly estimated step difficulties in the PCM and the calculated step difficulties in the five polytomous item explanatory models via correlations and graphical comparisons. As in Table 5, the three polytomous random item effects models revealed almost perfect agreement with the PCM in that their correlations were 0.99. This implies that item error terms worked well to enhance prediction of the polytomous item difficulties. The LPCM had a slightly higher correlation with the PCM than the MFRM (ρ = 0.91 for the MFRM, and ρ = 0.93 for the LPCM), although the LPCM fit worse than the MFRM.

Fig. 1 shows a graphical comparison of the estimated and calculated step difficulties between the models, which confirmed the

Table 4
Goodness-of-fit of models fitted to the Verbal Aggression data.

Model	q	DIC	LOOIC	WAIC
PCM	49	12312.49	12324.91	12320.29
MFRM	30	12448.42	12456.71	12452.81
MFRM + OIE	31	12312.39	12322.23	12317.86
LPCM	11	12454.69	12464.35	12460.36
LPCM + MISE	14	12299.66	12308.67	12304.47
LPCM + UISE	12	12307.09	12315.88	12311.65

Note: q = The number of estimated parameters.

Table 5
Correlations of the estimated and calculated step difficulties between models fitted to the Verbal Aggression data.

Model	PCM	MFRM	MFRM + OIE	LPCM	LPCM + MISE	LPCM + UISE
PCM	1	0.91	0.99	0.93	0.99	0.99
MFRM		1	0.94	0.96	0.93	0.94
MFRM + OIE			1	0.94	0.99	1
LPCM				1	0.95	0.96
LPCM + MISE					1	1
LPCM + UISE						1

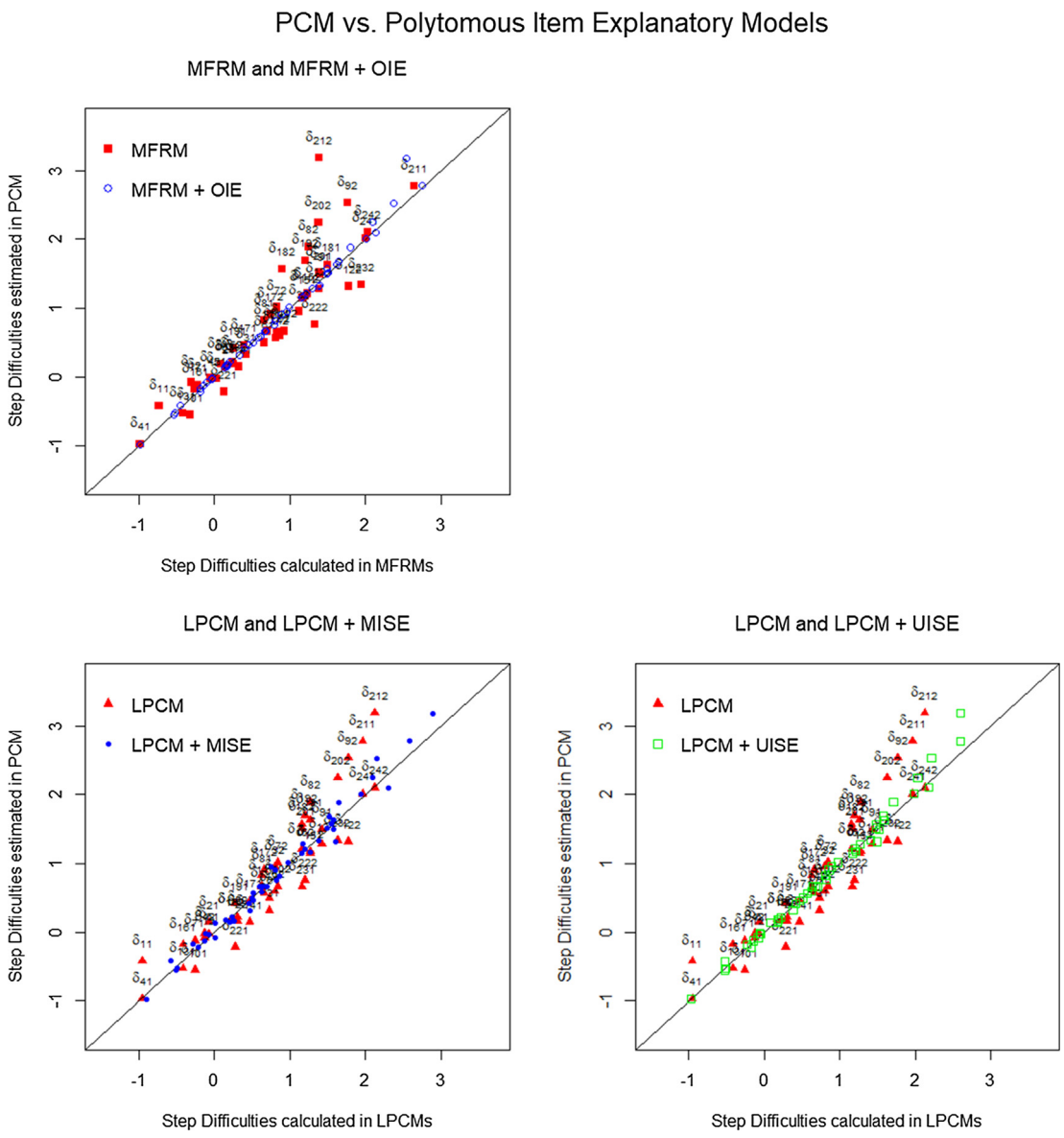

Fig. 1. Graphical agreement between models fitted to the Verbal Aggression data.

correlation analysis results. In three panels, the step difficulty points for the MFRM + OIE (blue circles), the LPCM + MISE (blue dots), and the LPCM + UISE (green squares) were closer than the two polytomous item explanatory models without random item errors to the 45-degree line indicating perfect alignment. Compared to the LPCM (red triangles) in the bottom panels, the MFRM (red squares) in the top panel had similar dispersion of the points but a few points were farther from the line. In particular, in both the MFRM and the LPCM, disagreements with the PCM were larger at the extremes (e.g., over ± 2). The second step difficulty of item 21 were located farthest from the 45-degree line in both models. Although it was the largest in the PCM, the polytomous item explanatory models without random item errors couldn't reconstruct it well. Those models estimate the item property effects for a set of items sharing the common item property, not for an individual item, so that there can be a discrepancy in the step difficulties between the individually estimated values and the calculated values by the estimated item property effects. However, the proposed polytomous random item effects models can compensate for the discrepancy by accounting for unexplained residual variation in the step difficulties even at the extremes.

Remember that the three categorical item properties (Behavior Mode, Situation Type, and Behavior Type) are from the item design factors of the Verbal Aggression items. To focus on methodological and practical differences between the item location explanatory MFRM and the step difficulty explanatory LPCM approaches, the effects of the three item properties on the overall item difficulties or the step difficulties were interpreted in each approach.

Table 6 presents the results of the fixed and random effects in the two MFRM-based polytomous item explanatory models. The estimated person variance (σ_0^2) was 0.95^2 for the MFRM and 0.97^2 for the MFRM + OIE. These variance estimates indicate that the shrinkage effect of the person variance estimate was small in the MFRM and the MFRM + OIE compensated for the shrinkage by enhancing prediction of the overall item difficulties.

In the MFRM + OIE, the overall item error variance was estimated as 0.32^2, which means significant residual variability in the overall item difficulties. R^2 was calculated as 0.85 using the estimated item property effects and the overall item error variance. As expected from the goodness-of-fit results, the three item properties had a high explanatory power on the overall item difficulties of the 24 Verbal Aggression items. In the MFRM, the standard

Table 6
Model parameter estimates and standard errors in the MFRM-based polytomous item explanatory models to the Verbal Aggression data.

Predictor	Parameter	MFRM		MFRM + OIE	
		Estimate	SE	Estimate	SE
Fixed effects					
Intercept	γ_0	1.58	0.08	1.69	0.18
Behavior Mode (Do)	γ_{Do}	0.43	0.04	0.49	0.15
Situation Type (Other-to-blame)	γ_{Other}	−0.82	0.04	−0.89	0.14
Behavior Type (Curse)	γ_{Curse}	−1.28	0.05	−1.38	0.17
Behavior Type (Scold)	γ_{Scold}	−0.63	0.05	−0.70	0.18
Step deviation parameters for each item	τ_{11}	−0.22	0.13	−0.17	0.12
	τ_{21}	0.00	0.13	−0.01	0.13
	τ_{31}	−0.35	0.13	−0.32	0.13
	τ_{41}	−0.47	0.12	−0.48	0.13
	τ_{51}	−0.11	0.12	−0.12	0.13
	τ_{61}	−0.11	0.13	−0.05	0.14
	τ_{71}	−0.52	0.12	−0.56	0.12
	τ_{81}	−0.29	0.13	−0.50	0.14
	τ_{91}	−0.18	0.15	−0.45	0.19
	τ_{101}	−0.62	0.12	−0.61	0.11
	τ_{111}	−0.25	0.13	−0.25	0.14
	τ_{121}	−0.19	0.15	−0.04	0.16
	τ_{131}	−0.34	0.12	−0.35	0.12
	τ_{141}	−0.25	0.13	−0.21	0.13
	τ_{151}	−0.03	0.14	−0.03	0.15
	τ_{161}	−0.17	0.13	−0.18	0.13
	τ_{171}	−0.17	0.13	−0.22	0.13
	τ_{181}	0.30	0.15	0.07	0.17
	τ_{191}	−0.48	0.12	−0.62	0.14
	τ_{201}	0.01	0.15	−0.30	0.18
	τ_{211}	0.63	0.21	0.10	0.27
	τ_{221}	−0.60	0.12	−0.49	0.13
	τ_{231}	−0.56	0.13	−0.36	0.13
	τ_{241}	−0.02	0.17	−0.06	0.21
Random effects					
Overall item error variance	σ^2_ε			0.32^2	0.07
Person variance	σ^2_θ	0.95^2	0.05	0.97^2	0.05

Note: Estimate = posterior mean; SE = posterior standard deviation (empirical standard error); an estimate and a standard error of standard deviation were obtained for the random effects from the Stan output.

errors of the item property effects were underestimated due to ignoring random residuals on the overall item difficulties. In the MFRM + OIE, however, they were properly estimated considering the random residuals, and all predictors of the item property effects including the intercept were statistically significant at the 5% level. This was consistent with such a high R^2. The item property effects in the MFRM + OIE are interpreted below (but the step deviation parameters τ are not of interest for interpretation in the item location explanatory MFRM approach).

For the Behavior Mode, holding other properties constant, the Doing mode made the overall item difficulty of an item 0.49 logits more difficult than the Wanting mode (γ_{Do} = 0.49, p = 0.001). In the Situation Type, the Other-to-blame situation made the overall item difficulty of an item 0.89 logits easier than the Self-to-blame situation (γ_{Other} = -0.89, p < .001), keeping other properties constant. For the Behavior Type, holding other properties constant, compared to Shouting, Cursing made the overall item difficulty of an item 1.38 logits easier (γ_{Curse} = -1.38, p < .001), and Scolding made the overall item difficulty of an item 0.70 logits easier (γ_{Scold} = -0.70, p < .001). These results are similar to ones reported earlier [15].

Table 7 shows the results for the fixed and random effects parameters in the three LPCM-based polytomous item explanatory models. The estimated person variance (σ^2_θ) was 0.95^2 for the LPCM, 0.97^2 for the LPCM + MISE, and 0.97^2 for the LPCM + UISE. The shrinkage effect of the person variance estimate was small in the LPCM and both the LPCM + MISE and the LPCM + UISE compensated for the shrinkage by enhancing prediction of the step difficulties, as found in the item location explanatory MFRM approach. Comparing the two LPCMs with polytomous random item errors, the LPCM + UISE performed as well as the LPCM + MISE in reconstructing the step difficulties and in compensating for the person variance shrinkage, however, the LPCM + MISE performed better than the LPCM + UISE in fitting to the Verbal Aggression data. This implies that both multivariate and univariate item-step random errors could account for unexplained residual variation in the step difficulties comparably well, but item-step random errors worked better for the Verbal Aggression data when the ordinal scale structure was considered. Thus, the results of the LPCM + MISE were examined and interpreted step-specifically for each step.

From the results of the LPCM + MISE, the item-step error variance was estimated as 0.33^2 for the first step and 0.31^2 for the second step, and their correlation was estimated as 0.82. Since the item-step error standard deviation for each step was significant at the 5% level, there was significant residual variability in the step difficulties for each step. Using the estimated step specific item property effects and the step specific item-step random error variances, R^2 was calculated as 0.87 for the first step and 0.86 for the second step. These R^2 values were high as in the MFRM + OIE. Compared to the underestimated standard errors in the LPCM, the LPCM + MISE considered random residuals on the step difficulties and provided the corrected standard errors. All predictors of the step specific item property effects including the step intercepts were statistically significant at the 5% level. This confirmed the high R^2 for both steps.

The effects of the three item properties on the step difficulties of the first step were interpreted as follows. When a person answered “perhaps” rather than “no”, keeping other properties constant, the

Table 7
Model parameter estimates and standard errors in the LPCM-based polytomous item explanatory models to the Verbal Aggression data.

Predictor	Parameter	LPCM		LPCM + MISE		LPCM + UISE	
		Estimate	SE	Estimate	SE	Estimate	SE
Fixed effects							
Intercept	ω_{01}	1.43	0.08	1.45	0.18	1.45	0.17
	ω_{02}	1.77	0.12	1.96	0.19	1.90	0.20
Behavior Mode (Do)	ω_{Do1}	0.54	0.06	0.53	0.14	0.56	0.14
	ω_{Do2}	0.36	0.07	0.44	0.15	0.42	0.16
Situation Type (Other-to-blame)	ω_{Other1}	−0.70	0.06	−0.71	0.15	−0.70	0.15
	ω_{Other2}	−0.97	0.07	−1.09	0.16	−1.06	0.16
Behavior Type (Curse)	ω_{Curse1}	−1.68	0.08	−1.71	0.17	−1.71	0.19
	ω_{Curse2}	−0.93	0.10	−1.11	0.18	−1.06	0.20
Behavior Type (Scold)	ω_{Scold1}	−0.80	0.07	−0.83	0.19	−0.84	0.18
	ω_{Scold2}	−0.50	0.10	−0.61	0.19	−0.56	0.20
Random effects							
Correlation between the steps	$\rho_{\xi 12}$			0.82	0.16		
Item-Step error variance (1st step)	$\sigma^2_{\xi 1}$			0.33^2	0.07		
Item-Step error variance (2nd step)	$\sigma^2_{\xi 2}$			0.31^2	0.08		
Univariate Item-Step error variance	σ^2_{ε}					0.33^2	0.05
Person variance	σ^2_{θ}	0.95^2	0.05	0.97^2	0.05	0.97^2	0.05

Note: Estimate = posterior mean; SE = posterior standard deviation (empirical standard error); an estimate and a standard error of standard deviation were obtained for the random effects from the Stan output.

Doing mode made the items 0.53 logits more difficult than the Wanting mode (ω_{Do1} = 0.53, p < .001), the Other-to-blame situation made the items 0.71 logits easier than the Self-to-blame situation (ω_{Other1} = -0.71, p < .001), and compared to Shouting, Cursing made the items 1.71 logits easier (ω_{Curse1} = -1.71, p < .001) and Scolding made the items 0.83 logits easier (ω_{Scold1} = -0.83, p < .001).

For the step difficulties of the second step, holding other properties constant, the Doing mode made the items 0.44 logits more difficult than the Wanting mode (ω_{Do2} = 0.44, p < .001), the Other-to-blame situation made the items 1.09 logits easier than the Self-to-blame situation (ω_{Other2} = -1.09, p < .001), and compared to Shouting, Cursing made the items 1.11 logits easier (ω_{Curse2} = -1.11, p < .001) and Scolding made the items 0.61 logits easier (ω_{Scold2} = -0.61, p < .001), as a person answered "yes" rather than "perhaps". These results are similar to ones reported earlier [15].

In practice, we could figure out how the three item design factors (Behavior Mode, Situation Type, and Behavior Type) explain and/or predict the overall item difficulties in the MFRM + OIE as well as the step difficulties in the LPCM + MISE. This empirical application demonstrated practical differences of the proposed polytomous item explanatory models with random item errors in interpreting the effects of the item properties on the overall item difficulties or the step difficulties of the Verbal Aggression items. Therefore, in addition to the methodological and conceptual differences in the two different polytomous item explanatory approaches and the three types of polytomous random item errors, the proposed models are practically different.

5. Conclusion and discussion

This paper has investigated how to extend the LLTM + ε approach to polytomous data. Considering the uncertainty in explanation and/or the random nature of item parameters, the concepts and types of polytomous random item effects were examined and then they were incorporated into the existing polytomous item explanatory models, the item location explanatory MFRM and the step difficulty explanatory LPCM. Through the two modeling steps of polytomous extensions of the LLTM + ε, the three polytomous item explanatory models with random item errors were proposed: the MFRM + OIE, the LPCM + MISE, and the LPCM + UISE. Using a Bayesian inference method, an empirical study was conducted to demonstrate practical implications and applications of the proposed models to the Verbal Aggression data.

From the empirical findings, the proposed polytomous item explanatory models with random item errors performed better in fitting the data and also better in reconstructing the step difficulties than the existing polytomous item explanatory models without random item errors. The results demonstrated methodological and practical differences of the proposed models in interpreting the item property effects in each of the two polytomous item explanatory approaches, the item location explanatory MFRM and the step difficulty explanatory LPCM approaches. In sum, The MFRM + OIE explains the overall item difficulties by the item property effects and compensates for the discrepancy by accounting for unexplained residuals regardless of the ordinal scale structure, so that it can enhance to predict the overall item difficulties. The LPCM + MISE or the LPCM + UISE explain the step difficulties by the step specific item property effects and compensate for the discrepancy by accounting for unexplained residuals with or without consideration for the ordinal scale structure, so that they can enhance to predict the step difficulties.

In educational and psychological measurement research, random item effects models are a rather new approach and most applications of them have been limited to dichotomous items (e.g., [9,16,26,28,31,34,35]). However, little is known about random item effects for polytomous items. This paper tried to shed light on the methodological advantages and the practical implications of polytomous random item effects models in the context of item explanatory modeling for polytomous data.

Polytomous random item effects models are emerging and promising. In addition to the findings from the empirical example, we emphasize potential uses of them as methodological foundations in IRT modeling. First, although we have focused on item explanatory modeling (i.e., polytomous item explanatory models with random item errors), the proposed models could be combined with other explanatory models such as person explanatory and doubly explanatory models (see [15]). Second, if item properties and individual persons interact, the proposed models could be extended to multidimensional versions. This is analogous to the Random-Weights Linear Logistic Test Model (RW-LLTM; [60]) for polytomous items. If item properties and fixed groups of persons interact, the proposed models could be extended with differential facet functioning (DFF; see [15]). Third, the proposed models are item explanatory models as well as random item effects models by nature so that they could boost potential uses of them for polytomous data. For example,

if the items are nested within item groups, the proposed models could be extended to hierarchical polytomous item structure models with random residuals (see [10]). If finite mixtures of underlying distributions of the person and/or item population are assumed, the proposed models could be combined with latent class models (see [16,20,25]). Fourth, the proposed models could be used for making polytomous item families and item banks. This stresses the generalizability potential and the random sampling interpretation in computer adaptive testing, automatic item generation, and generalizability in item response modeling (see [7,12,34,36,62]).

Acknowledgments

The authors would like to thank Sophia Rabe-Hesketh for her careful comments on model specification and thank anonymous reviewers for their helpful comments on an earlier draft.

Declaration of conflicting interests

The authors declared no potential conflicts of interests with respect to the research, authorship and/or publication of this article.

Funding

Development of Stan codes for estimating the proposed models in this paper was supported in part by Grant R305D140059 from the Institute of Education Sciences (IES), U.S. Department of Education. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

Appendix

A.1. Priors of model parameters for polytomous item explanatory models with random item errors

Priors of model parameters for the MFRM + OIE

$\theta_p \sim N(0,\ \sigma_\theta^2)$,
$\sigma_\theta \sim Cauchy(0,\ 5)$, truncated at 0,
$\gamma_k \sim N\left(0,\ 10^2\right)$,
$\tau_{im} \sim N\left(0,\ 3^2\right)$,
$\in_i \sim N(0,\ \sigma_\in^2)$,
$\sigma_\in \sim Cauchy(0,\ 5)$, truncated at 0.

Priors of model parameters for the LPCM + MISE

$\theta_p \sim N(0,\ \sigma_\theta^2)$,
$\sigma_\theta \sim Cauchy(0,\ 5)$, truncated at 0,
$\omega_{km} \sim N\left(0, 10^2\right)$ for $m \in 1:M$,
$\boldsymbol{\xi_i} \sim MVN_m(\mathbf{0},\ \boldsymbol{\Sigma})$, where $\boldsymbol{\xi_i} = (\xi_{i1},\ \xi_{i2},\ \ldots,\ \xi_{im})'$,
$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma_\xi}) \times \boldsymbol{\Omega} \times \mathrm{diag}(\boldsymbol{\sigma_\xi})$, where $\boldsymbol{\sigma_\xi} = (\sigma_{\xi 1}, \sigma_{\xi 2},\ \ldots, \sigma_{\xi m})'$,
$\sigma_{\xi m} \sim Cauchy(0,\ \ 2.5)$ for $m \in 1:M$, truncated at 0,
$\boldsymbol{\Omega} \sim LKJ(1)$.

Priors of model parameters for the LPCM + UISE

$\theta_p \sim N(0,\ \ \sigma_\theta^2)$,
$\sigma_\theta \sim Cauchy(0,\ 5)$, truncated at 0,
$\omega_{km} \sim N\left(0,\ 10^2\right)$ for $m \in 1:M$,
$\varepsilon_{im} \sim N(0,\ \sigma_\varepsilon^2)$,
$\sigma_\varepsilon \sim Cauchy(0,\ 5)$, truncated at 0.

The prior for the person ability parameter θ_p in all models is a normal distribution with zero mean. The hyperprior for the standard deviation is a Cauchy distribution truncated to the positive real values, i.e., a half-Cauchy distribution, with a location of zero and a scale of 5. This half-Cauchy is a weakly informative prior that provides a large mass near the boundary and also flattens out at the boundary, regarded as a good default choice for scale parameters [41,65]. In addition, the priors for the fixed item effects parameters, the item property effect parameters γ_k and ω_{km} and step deviation parameter τ_{im}, have weakly informative priors. Each parameter follows a normal distribution with a zero mean (for parameter identifiability) and a fairly large standard deviation (for flatter priors). For the random item effects parameters ϵ_i, ξ_{im}, and ε_{im}, each of the priors is a normal distribution with a zero mean as a usual error, and a half-Cauchy distribution is specified as a weakly informative hyperprior for the standard deviation of each prior.

In particular, the multivariate item-step random error vector $\boldsymbol{\xi_i}(\xi_{i1}, \xi_{i2}, \cdots, \xi_{im})'$ in the LPCM + MISE follows a multivariate normal (MVN) distribution with a zero mean vector $\mathbf{0}$, and the prior for the variance–covariance matrix $\boldsymbol{\Sigma}(m \times m)$ is decomposed into a scale or standard deviation vector $\boldsymbol{\sigma_\xi}$ $(\sigma_{\xi 1}, \sigma_{\xi 2}, \cdots, \sigma_{\xi m})'$ and a correlation matrix $\boldsymbol{\Omega}$. The components of the scale vector $\sigma_{\xi m}$ have a weakly informative hyperprior in each step m that follows a half-Cauchy distribution with a small scale of 2.5, as recommend by the Stan manual [65]. It is also recommended that an LKJ [42] prior is placed on the correlation matrix $\boldsymbol{\Omega}$, and the LKJ correlation distribution with a shape of one is weakly informative due to reducing to the identity distribution. Since it is computationally more efficient and arithmetically stable to use a Cholesky factor of the variance–covariance matrix $\boldsymbol{\Sigma}$ [65], the Cholesky decomposition of $\boldsymbol{\Sigma}$ is also implemented. Lastly, although the fixed and random item effects are step specific in the LPCM + MISE, a univariate normal distribution is used as the prior for the item property effect parameter ω_{km} because it produces the same density much more efficiently to the weakly informative multivariate normal prior with no correlation information in Stan [65].

These priors are recommended by the Stan manual and other Bayesian IRT studies (e.g., [14,24,38,41,45]).

A.2. Stan codes for polytomous item explanatory models with random item errors

Stan codes for the MFRM + OIE

```
functions {
  real pcm(int y, real theta, vector delta) {
    vector[rows(delta) + 1] unsummed;
    vector[rows(delta) + 1] probs;
    unsummed = append_row(rep_vector(0.0, 1), theta - delta);
    probs = softmax(cumulative_sum(unsummed));
    return categorical_lpmf(y + 1 | probs);
}}
data {
  int < lower = 1 > P; // number of persons
  int < lower = 1 > I; // number of items/questions
  int < lower = 1 > M; // number of steps per item (same for all
items)
  int < lower = 1 > N; // number of observations (P*I; as if they
were in a vector)
  int < lower = 1, upper = P > pp[N]; // person p for observation
n
  int < lower = 1, upper = I > ii[N]; // item i for observation n
```

```
    int < lower = 0 > y[N]; // reponse/score for observation n: y = 0, 1 ... m_i
    int < lower = 1 > K; // number of item predictors
    matrix[I,K] Q; // matrix of item predictors
  }
  parameters {
    vector[K] betaip; // betaip: item property effects on overall item difficulty
    vector[M-1] tau_free[I]; // tau: step deviation
    vector[I] ierror;
    real < lower = 0 > sigmai; // sigmai: sd of item errors
    vector[P] theta; // theta: person ability
    real < lower = 0 > sigmap; // sigmap: sd of person abilities
  }
  transformed parameters {
    vector[I] beta; // beta: overall item difficulty
    vector[M] tau[I]; // tau: step deviation
    vector[M] delta[I]; // delta: step difficulty
    beta = Q*betaip + ierror;
    for(i in 1:I) for(m in 1:M-1)
      tau[i,m] = tau_free[i,m];
    for(i in 1:I)
      tau[i,M] = -1*sum(tau_free[i,1:M-1]);
    for(i in 1:I) for(m in 1:M)
      delta[i,m] = beta[i] + tau[i,m];
  }
  model {// Case constraint (mean of thetas equals to zero)
    target += normal_lpdf(theta | 0, sigmap);
    sigmap ~ cauchy(0, 5) T[0, ]; // hyperprior for theta's sd
    target += normal_lpdf(betaip | 0, 10);
    target += normal_lpdf(ierror | 0, sigmai);
    sigmai ~ cauchy(0, 5) T[0, ];
    for(i in 1:I)
      target += normal_lpdf(tau[i] | 0, 3);
    for (n in 1:N)
      target += pcm(y[n], theta[pp[n]], delta[ii[n]]); // likelihood
  }
  generated quantities {
    vector[N] log_lik;
    real deviance;
    for (n in 1:N)
      log_lik[n] = pcm(y[n], theta[pp[n]], delta[ii[n]]);
    deviance = sum(-2*log_lik);
  }
```

Stan codes for the LPCM + MISE

```
  functions {
    real pcm(int y, real theta, vector delta) {
    vector[rows(delta) + 1] unsummed;
    vector[rows(delta) + 1] probs;
    unsummed = append_row(rep_vector(0.0, 1), theta - delta);
    probs = softmax(cumulative_sum(unsummed));
    return categorical_lpmf(y + 1 | probs);
  }}
  data {// vectorized version
    int < lower = 1 > P; // number of persons
    int < lower = 1 > I; // number of items/questions
    int < lower = 1 > M; // number of steps per item (same for all items)
    int < lower = 1 > N; // number of observations (P*I; as if they were in a vector)
    int < lower = 1, upper = P > pp[N]; // person p for observation n
    int < lower = 1, upper = I > ii[N]; // item i for observation n
    int < lower = 0 > y[N]; // reponse/score for observation n: y = 0, 1 ... m_i
    int < lower = 1 > K; // number of item predictors
    matrix[I,K] Q; // matrix of item predictors
  }
  parameters {
    matrix[K,M] deltaip; // deltaip: item property effects on step difficulty
    matrix[I,M] sserror;
    corr_matrix[M] Omega;
    vector < lower = 0 > [M] sigmad; // sd of step difficulties for each step
    vector[P] theta; // theta: person ability
    real < lower = 0 > sigmap; // sigmap: sd of person abilities
  }
  transformed parameters {
    matrix[I,M] delta; // delta: calculated step difficulty
    cov_matrix[M] Vard;
    delta = Q*deltaip + sserror;
    Vard = quad_form_diag(Omega, sigmad);
  }
    model {// Case constraint (mean of thetas equals to zero)
    matrix[M,M] chol_Sigma;
    target += normal_lpdf(theta | 0, sigmap);
    sigmap ~ cauchy(0, 5) T[0, ]; // hyperprior for theta's sd
    target += normal_lpdf(to_vector(deltaip) | 0, 10);
    chol_Sigma = cholesky_decompose(Vard);
    target += lkj_corr_lpdf(Omega | 1);
    target += cauchy_lpdf(sigmad | 0, 2.5);
    for (i in 1:I)
      target += multi_normal_cholesky_lpdf(to_vector(sserror[i,]) | rep_vector(0, M), chol_Sigma);
  for (n in 1:N)
    target += pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]])); // likelihood
  }
  generated quantities {
    vector[N] log_lik;
    real deviance;
    for (n in 1:N)
      log_lik[n] = pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]]));
    deviance = sum(-2*log_lik); //the sum of its unit deviances
  }
```

Stan codes for the LPCM + UISE

```
  functions {
    real pcm(int y, real theta, vector delta) {
      vector[rows(delta) + 1] unsummed;
      vector[rows(delta) + 1] probs;
      unsummed = append_row(rep_vector(0.0, 1), theta - delta);
      probs = softmax(cumulative_sum(unsummed));
      return categorical_lpmf(y + 1 | probs);
  }}
  data {
    int < lower = 1 > P; // number of persons
    int < lower = 1 > I; // number of items/questions
    int < lower = 1 > M; // number of steps per item (same for all items)
    int < lower = 1 > N; // number of observations (P*I; as if they were in a vector)
    int < lower = 1, upper = P > pp[N]; // person p for observation n
    int < lower = 1, upper = I > ii[N]; // item i for observation n
    int < lower = 0 > y[N]; // reponse/score for observation n: y = 0, 1 ... m_i
    int < lower = 1 > K; // number of item predictors
    matrix[I,K] Q; // matrix of item predictors
  }
```

```
parameters {
  matrix[K,M] deltaip; // deltaip: item property effects on step difficulty
  matrix[I,M] iserror;
  real < lower = 0 > sigmais; // sigmais: sd of item errors
  vector[P] theta; // theta: person ability
  real < lower = 0 > sigmap; // sigmap: sd of person abilities
}
transformed parameters {
  matrix[I,M] delta; // delta: calculated step difficulty
  delta = Q*deltaip + iserror;
}
model {// Case constraint (mean of thetas equals to zero)
  target += normal_lpdf(theta | 0, sigmap);
  sigmap ~ cauchy(0, 5) T[0, ]; // hyperprior for theta's sd
   target += normal_lpdf(to_vector(deltaip) | 0, 10);
  target += normal_lpdf(to_vector(iserror) | 0, sigmais);
  sigmais ~ cauchy(0, 5) T[0, ];
for (n in 1:N)
  target += pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]])); // likelihood
}
generated quantities {
  vector[N] log_lik;
  real deviance;
for (n in 1:N)
  log_lik[n] = pcm(y[n], theta[pp[n]], to_vector(delta[ii[n]]));
  deviance = sum(-2*log_lik);
}
```

References

[1] R.J. Adams M. Wu M. Wilson, ConQuest 3.0 [computer program]. ACER, Hawthorn, Australia, 2012.
[2] A.J. Ames, K. Samonte, Using SAS PROC MCMC for item response theory models, Educ. Psychol. Measur. 75 (4) (2015) 585–609.
[3] J.A. Anderson, Regression and ordered categorical variables (with discussion), J. R. Stat. Soc. 46 (1984) 1–30.
[4] D. Andrich, A rating formulation for ordered response categories, Psychometrika 43 (4) (1978) 561–573.
[5] R. Bellio, C. Varin, A pairwise likelihood approach to generalized linear models with crossed random effects, Stat. Modell. 5 (3) (2005) 217–227.
[6] P. Black, D. Wiliam, Assessment and classroom learning, Assess. Educ.: Principles, Policy, Pract. 5 (1) (1998) 7–74.
[7] D.C. Briggs, M. Wilson, Generalizability in item response modeling, J. Educ. Meas. 44 (2) (2007) 131–155.
[8] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, A. Riddell, Stan: A probabilistic programming language, J. Stat. Softw. 76 (1) (2017) 1–32.
[9] S.J. Cho, S. Rabe-Hesketh, Alternating imputation posterior estimation of models with crossed random effects, Comput. Stat. Data Anal. 55 (1) (2011) 12–25.
[10] S.J. Cho, P. De Boeck, S. Embretson, S. Rabe-Hesketh, Additive multilevel item structure models with random residuals: item modeling for explanation and item generation, Psychometrika 79 (1) (2014) 84–104.
[11] S.-J. Cho, I. Partchev, P. De Boeck, Parameter estimation of multiple item response profiles model, Br. J. Math. Stat. Psychol. 65 (2012) 438–466.
[12] J.I. Choi, Advances in combining Generalizability Theory and Item Response Theory (Unpublished doctoral dissertation), University of California, Berkeley, 2013.
[13] F. Cristante, E. Robusto, Assessing change with the extended logistic model, Br. J. Math. Stat. Psychol. 60 (2) (2007) 367–375.
[14] S.M. Curtis, BUGS code for item response theory, J. Stat. Softw. 36 (1) (2010) 1–34.
[15] P. De Boeck, M. Wilson (Eds.), Explanatory item response models: A generalized linear and nonlinear approach, Springer-Verlag, New York, 2004.
[16] P. De Boeck, Random item IRT models, Psychometrika 73 (4) (2008) 533–559.
[17] P. De Boeck, M. Bakker, R. Zwitser, M. Nivard, A. Hofman, F. Tuerlinckx, I. Partchev, The estimation of item response models with the lmer function from the lme4 package in R, J. Stat. Softw. 39 (12) (2011) 1–28.
[18] S.E. Embretson, S.P. Reise, Item response theory for psychologists, LEA, Mahwah, NJ, 2000.
[19] S. Embretson, J. Gorin, Improving construct validity with cognitive psychology principles, J. Educ. Meas. 38 (4) (2001) 343–368.
[20] S. Fieuws, B. Spiessens, K. Draney, Mixture models, in: P. De Boeck, M. Wilson (Eds.), Explanatory item response models: A generalized linear and nonlinear approach, Springer-Verlag, New York, 2004, pp. 317–340.
[21] G.H. Fischer, The linear logistic test model as an instrument in educational research, Acta Psychol. 37 (1973) 359–374.
[22] G.H. Fischer, P. Parzer, An extension of the rating scale model with an application to the measurement of change, Psychometrika 56 (4) (1991) 637–651.
[23] G.H. Fischer, I. Ponocny, An extension of the partial credit model with an application to the measurement of change, Psychometrika 59 (2) (1994) 177–192.
[24] J.P. Fox, Bayesian item response modeling: Theory and applications, Springer, New York, 2010.
[25] S. Frederickx, F. Tuerlinckx, P. De Boeck, D. Magis, RIM: A random item mixture model to detect differential item functioning, J. Educ. Meas. 47 (4) (2010) 432–457.
[26] D.C. Furr, Bayesian and frequentist cross-validation methods for explanatory item response models (Unpublished doctoral dissertation), University of California, Berkeley, 2017.
[27] A. Gelman, D.B. Rubin, Inference from iterative simulation using multiple sequences, Stat. Sci. 7 (4) (1992) 457–472.
[28] C.A. Glas, W.J. van der Linden, Computerized adaptive testing with item cloning, Appl. Psychol. Meas. 27 (4) (2003) 247–261.
[29] C.A.W. Glas, N.D. Verhelst, Extensions of the partial credit model, Psychometrika 54 (4) (1989) 635–659.
[30] S.M. Haley, P. Ni, R.K. Hambleton, M.D. Slavin, A.M. Jette, Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank, J. Clin. Epidemiol. 59 (11) (2006) 1174–1182.
[31] J. Hartig, A. Frey, G. Nold, E. Klieme, An application of explanatory item response modeling for model-based proficiency scaling, Educ. Psychol. Measur. 72 (4) (2012) 665–686.
[32] J. Hartzel, A. Agresti, B. Caffo, Multinomial logit random effects models, Stat. Modell. 1 (2) (2001) 81–102.
[33] W. Hively, Introduction to domain-referenced testing, Educ. Technol. 14 (6) (1974) 5–10.
[34] H. Holling, J.P. Bertling, N. Zeuch, Automatic item generation of probability word problems, Stud. Educ. Eval. 35 (2–3) (2009) 71–76.
[35] R. Janssen, J. Schepers, D. Peres, Models with item and item group predictors, in: P. De Boeck, M. Wilson (Eds.), Explanatory item response models: A generalized linear and nonlinear approach, Springer-Verlag, New York, 2004, pp. 189–212.
[36] M.S. Johnson, S. Sinharay, Calibration of polytomous item families using Bayesian hierarchical modeling, Appl. Psychol. Meas. 29 (5) (2005) 369–400.
[37] T.R. Johnson, Discrete choice models for ordinal response variables: A generalization of the stereotype model, Psychometrika 72 (2007) 489–504.
[38] T. Kang, A.S. Cohen, H.J. Sung, Model selection indices for polytomous items, Appl. Psychol. Meas. 33 (7) (2009) 499–518.
[39] J. Kim, Extensions and Applications of Item Explanatory Models to Polytomous Data in Item Response Theory (Unpublished doctoral dissertation), University of California, Berkeley, 2018.
[40] K.D. Kubinger, Applications of the linear logistic test model in psychometric research, Educ. Psychol. Measur. 69 (2) (2009) 232–244.
[41] Q.N. Lathrop, Y. Cheng, Item cloning variation and the impact on the parameters of response models, Psychometrika 82 (1) (2017) 245–263.
[42] D. Lewandowski, D. Kurowicka, H. Joe, Generating random correlation matrices based on vines and extended onion method, J. Multivariate Anal. 100 (9) (2009) 1989–2001.
[43] J.M. Linacre, Multi-facet Rasch measurement, MESA Press, Chicago, 1989.
[44] D.J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility, Stat. Comput. 10 (4) (2000) 325–337.
[45] Y. Luo, H. Jiao, Using the Stan program for Bayesian item response theory, Educ. Psychol. Measur. 78 (3) (2018) 384–408.
[46] G.N. Masters, A Rasch model for partial credit scoring, Psychometrika 47 (2) (1982) 149–174.
[47] G.N. Masters, B.D. Wright, The essential process in a family of measurement models, Psychometrika 49 (4) (1984) 529–544.
[48] G.N. Masters, B.D. Wright, The partial credit model, in: Handbook of modern item response theory, Springer, New York, 1997, pp. 101–121.
[49] G.J. Mellenbergh, Conceptual notes on models for discrete polytomous item responses, Appl. Psychol. Meas. 19 (1) (1995) 91–100.
[50] R.J. Mislevy, Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters, Appl. Psychol. Meas. 12 (3) (1988) 281–296.
[51] I.W. Molenaar, Item steps (Heymans Bulletin 83-630-EX), University of Groningen, Department of Statistics and Measurement Theory, Groningen, The Netherlands, 1983.
[52] I.W. Molenaar, Nonparametric models for polytomous responses, in: Handbook of modern item response theory, Springer, New York, 1997, pp. 369–380.
[53] C.C. Monnahan, J.T. Thorson, T.A. Branch, Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo, Methods Ecol. Evol. 8 (3) (2017) 339–348.
[54] E. Muraki, R.D. Bock, PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales (computer software), Scientific Software International, Chicago, 1997.

[55] D.A. Pastor, B.G. Dodd, H.H. Chang, A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model, Appl. Psychol. Meas. 26 (2) (2002) 147–163.
[56] R.D. Penfield, Applying Bayesian item selection approaches to adaptive tests using polytomous items, Appl. Measur. Educ. 19 (1) (2006) 1–20.
[57] J. Piironen, A. Vehtari, Comparison of Bayesian predictive methods for model selection, Stat. Comput. 27 (3) (2017) 711–735.
[58] H. Poinstingl, The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test, Psychol. Sci. Quaterly 51 (2) (2009) 123–134.
[59] G. Rasch, Probabilistic models for some intelligence and attainment tests, The Danish Institute of Educational Research, Copenhagen, 1960. (Expanded edition, 1980. Chicago: The University of Chicago Press.).
[60] F. Rijmen, P. De Boeck, The random weights linear logistic test model, Appl. Psychol. Meas. 26 (3) (2002) 271–285.
[61] J. Rost, The growing family of Rasch models, in: A. Boomsma, M. van Duijin, T. Snijders (Eds.), Essays on item response theory, Springer, New York, 2001, pp. 25–42.
[62] S. Sinharay, M.S. Johnson, D.M. Williamson, Calibrating item families and summarizing the results using family expected response functions, J. Educ. Behav. Sci. 28 (2003) 295–313.
[63] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. van der Linde, Bayesian measures of model complexity and fit, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 64 (4) (2002) 583–639.
[64] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. van der Linde, The deviance information criterion: 12 years on, J. Roy. Stat. Soc. B (Statistical Methodology) 76 (3) (2014) 485–493.
[65] Stan Development Team, Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0, 2017. http://mc-stan.org.
[66] Stan Development Team, RStan: the R interface to Stan. R package version 2.17.3, 2018. http://mc-stan.org.
[67] StataCorp, Stata Base Reference Manual: Release 14, Stata Press, College Station, TX, 2015.
[68] F. Tuerlinckx, W.C. Wang, Models for polytomous data, in: P. De Boeck, M. Wilson (Eds.), Explanatory item response models: A generalized linear and nonlinear approach, Springer-Verlag, New York, 2004, pp. 75–109.
[69] W. Van den Noortgate, P. De Boeck, M. Meulders, Cross-classification multilevel logistic models in psychometrics, J. Educ. Behav. Stat. 28 (4) (2003) 369–386.
[70] K. Vansteelandt, Formal models for contextualized personality psychology. (Unpublished doctoral dissertation), K.U.Leuven, Belgium, 2000.
[71] A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, Stat. Comput. 27 (5) (2017) 1413–1432.
[72] A. Vehtari, A. Gelman, J. Gabry, Y. Yao, loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models. R package version 2.0.0, 2018.
[73] S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, J. Mach. Learn. Res. 11 (2010) 3571–3594.
[74] M.F. Zimowski, E. Muraki, R.J. Mislevy, R.D. Bock, BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software], Scientific Software International, Chicago, 1996.
[75] W.C. Wang, M. Wilson, Exploring local item dependence using a random-effects facet model, Applied Psychological Measurement 29 (4) (2005) 296–318.