International Journal of Assessment Tools in Education
2019, Vol. 6, No. 2, 259–278
https://dx.doi.org/10.21449/ijate.515085

Published at http://www.ijate.net http://dergipark.org.tr/ijate *Research Article*

Explanatory Item Response Models for Polytomous Item Responses

Luke Stanke [1*], **Okan Bulut** [2]

[1] Tessellation Minneapolis, MN, USA
[2] Centre for Research in Applied Measurement and Evaluation University of Alberta

ARTICLE HISTORY

Received: 19 January 2019
Revised: 08 May 2019
Accepted: 15 May 2019

KEYWORDS

Polytomous IRT,
Explanatory item response modeling,
Assessment,
Partial Credit Model

Abstract: Item response theory is a widely used framework for the design, scoring, and scaling of measurement instruments. Item response models are typically used for dichotomously scored questions that have only two score points (e.g., multiple-choice items). However, given the increasing use of instruments that include questions with multiple response categories, such as surveys, questionnaires, and psychological scales, polytomous item response models are becoming more utilized in education and psychology. This study aims to demonstrate the application of explanatory item response theory (IRT) models to polytomous item responses in order to explain common variability in item clusters, person groups, and interactions between item clusters and person groups. Explanatory forms of several IRT models – such as Partial Credit Model and Rating Scale Model – are demonstrated and the estimation procedures of these models are explained. Findings of this study suggest that explanatory IRT models can be more parsimonious than traditional IRT models for polytomous data when items and persons share common characteristics. Explanatory forms of the polytomous IRT models can provide more information about response patterns in item responses by estimating fewer item parameters.

1. INTRODUCTION

Item response theory (IRT) models have been widely used for the design, scoring, and scaling of educational and psychological assessments during the past three decades (Bond & Fox, 2001; Embretson & Reise, 2000; Lord, 1980; van der Linden & Hambleton, 1997; Wright & Masters, 1982). Dichotomous IRT models, such as the Rasch model (RM; Rasch, 1960/1980) and two-parameter logistic model (2PL; Birnbaum, 1968), have been more common in practice due to the popularity of standardized assessments with dichotomously scored multiple-choice items. However, today's educators desire to differentiate their students with more innovative assessment tools that consist of not only dichotomous items but also items with more than one score level (i.e., polytomous items). Similarly, many researchers prefer to use surveys, questionnaires, and scales with Likert-type items that often consist of multiple, ordered response categories (e.g., strongly disagree, disagree, agree, and strongly agree). To

CONTACT: Okan BULUT ✉ bulut@ualberta.ca Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB, T6G 2G5 Canada

ISSN-e: 2148-7456 /© IJATE 2019

accommodate such items, polytomous IRT models need to be utilized. Polytomous IRT models, including the Nominal Response Model (NRM; Bock, 1972), the Graded Response Model (GRM; Samejima, 1969), the Sequential Response Model (SRM; Tutz, 1990, 1991), the Rating Scale Model (RSM; Andrich, 1978), and the Partial Credit Model (PCM; Masters, 1982), can be used for items with either nominal or ordered response categories.

The traditional IRT models – regardless of number of response categories – can only provide direct information regarding respondents' trait levels in aptitude, achievement, cognitive abilities, and so on, as well as item information concerning the difficulty, discrimination, and fit of an item to the selected IRT model. Although researchers and practitioners often use these descriptive measures for making decisions regarding respondents and items, traditional IRT models are not able to identify any systematic effects that result from the design of a measurement instrument. That is, these models do not explain common variability across items or across respondents based on the design or theory behind the instrument. Measuring the commonality of responses is an important step in test development because it allows test developers to assess the degree which construct-relevant or construct-irrelevant features – including linguistic, communicative, cognitive, cultural, or physical features – are related to the construct being measured (AERA, APA, NCME, 2014). For example, consider a test that is taken by examinees who either are native speakers of English or speak English as a secondary language, but understanding the English language is not relevant to the target construct being measured. Under traditional IRT models, there would be no way of directly estimating the mean difference between primary and secondary English speakers' performances.

Expanding the same example, assume that this test assesses mathematical knowledge for middle-school students and researchers are interested in examining the potential effects of including graphics on test items to assist students in answering the items. The researchers can create two equivalent test forms where one test form contains images on half of the items, while the remaining items have no images. The second form contains the same items as the first form, but the presence or absence of images on items is the opposite of the first form. Using traditional IRT models, there would be no way of directly estimating the impact of using images on the difficulty levels of these test forms.

Information about the mean differences between primary and secondary English speakers or the impact of images in test items can be directly estimated using Explanatory Item Response Modeling (EIRM). De Boeck and Wilson (2004) introduced the EIRM framework for measuring common variability in item clusters, respondent groups, or the interactions between item clusters and respondent groups. Instead of estimating the descriptive effects of respondents' trait level or item difficulty, the explanatory item response models extract information from responses by including explanatory variables. Under the EIRM framework, traditional IRT models can be formulated as a subset of models that belong to a larger class of models – generalized linear mixed models (GLMMs). GLMMs can function as explanatory IRT models when the model includes an item covariate, a person covariate, or a person-by-item covariate (De Boeck & Wilson, 2004; Wilson, De Boeck, & Carstensen, 2008). The EIRM approach defines responses to items as repeated measures nested within each respondent in a multilevel framework. Within a multilevel model, the effects of explanatory variables can be estimated either as fixed or random effects. The linear logistic test model (LLTM; Fischer, 1973; De Boeck, 2008), the latent regression Rasch model (Zwinderman, 1991), and the latent regression LLTM are widely-used forms of explanatory IRT models (Desjardins & Bulut, 2018).

With EIRM, the object of measurement is typically not at the item or respondent levels, but a higher level to explain the relationship among the items or respondents. In the earlier example, explanatory IRT models can provide information to explain the mean differences between primary and secondary speakers of English, and help determine whether there is any impact of including images on the difficulty of items within the same model. Therefore, researchers can analyze item response data from tests using a perspective that goes beyond common practices in psychology and educational measurement (De Boeck & Wilson, 2004, p.7). The inclusion of explanatory variables in IRT models is typically based on a pre-defined theory. In the case of the example above, explanatory variables indicating English as a primary language (a person covariate) and presence of an image in an item (an item covariate) could be included as predictors in the same model.

1.1. Significance of Study

To date, EIRM has been mostly applied to either dichotomous data or pseudo-dichotomous data where polytomous response categories have been collapsed into binary categories through the selective grouping of ordered or nominal response categories (e.g., Bulut, Palma, Rodriguez, & Stanke, 2015; De Boeck & Partchev, 2012; Plieninger & Meiser, 2014; Prowker & Camilli, 2007; Scheiblechner, 2009; Verhelst & Verstralen, 2008). Despite more recent attempts that described how to estimate explanatory IRT models for items with ordered or nominal response categories (e.g., Jiao & Zhang, 2014; Wang & Liu, 2007; Tuerlinckx & Wang, 2004), the proposed models have been limited in terms of utilizing a familiar polytomous IRT model (e.g., GRM, PCM, and RM) within the EIRM framework. Also, these models mostly focused on the first threshold between item response categories as it is often interpreted as the difficulty of polytomous items. In this study, we aim to establish a basis for explanatory IRT models for polytomous item response data, not by formulating a new model, but elucidating the flexibility and usefulness of the existing polytomous explanatory IRT models. We used a real dataset to demonstrate the utility of the explanatory IRT models by examining the threshold parameters and model fit statistics. In addition, we described a new parameterization of the explanatory IRT models for polytomous response data that allows a straightforward estimation of these models in R (R Core Team, 2018).

1.2. Theoretical Background

1.2.1. *Explanatory Item Response Modeling*

Explanatory item response models can utilize IRT for both measurement and explanation purposes (De Boeck & Wilson, 2004). The main advantage of these models is the flexibility to analyze items and respondents, while simultaneously decomposing common variability across item- and respondent groups (Briggs, 2008). In addition, EIRM allows a theory to be directly imputed into IRT models. EIRM has been applied to a wide array of psychometric and measurement studies, including construct validity studies aiming to explain common variability in item parameters (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002; Embretson, 2006), latent growth modeling (Wilson, Zheng, & McGuire, 2012), local item dependence studies (Wang & Wilson, 2005), differential functioning (Luppescu, 2002; Williams & Beretvas, 2006; French & Finch, 2010), item parameter drift (Bulut et al., 2015), and contextual effect studies (Albano, 2013; Kan & Bulut, 2014; Kubinger, 2008).

Despite the increasing popularity of EIRM in educational and psychological settings, there are only a few instances of EIRM where researchers used explanatory variables to explain the item-level or person-level variation in polytomous response data. One of the very first attempts to study EIRM with polytomous data was Tuerlinckx and Wang's (2004) study where the authors fit a series of models to a verbal aggression dataset that consisted of 24 items with three response categories. The study examined model fit of five models: a RSM, an explanatory RSM with two person covariates, a PCM, an explanatory PCM with two person covariates, and an explanatory PCM with using five item covariates crossed with threshold parameters and two person covariates. The two polytomous explanatory models with only person covariates estimated two more parameters than their traditional counterparts. In both cases, the explanatory models with person characteristics fit better than the traditional polytomous IRT models. However, neither of these explanatory models could explain the location of thresholds or the distance between thresholds in the items.

Tuerlinckx and Wang's explanatory IRT model with both item and person covariates is the most interesting of the three polytomous explanatory models. This model estimates 13 parameters in total – 5 item explanatory variables for the first threshold, 5 item explanatory variables for the second threshold parameter, the two person covariates, and the variance component for the person trait level. Compared to the traditional PCM, the model estimates 36 fewer parameters. Unlike the first two explanatory models, this model uses covariates to explain the location of thresholds. Threshold locations for each item can be approximated by summing the coefficients where relevant item characteristics are present. Using AIC and BIC, this model did not fit as well compared to the other two explanatory models. While both the first and second threshold locations are estimated using item covariates, the parameterization of this model does not make it easy to explain the distance between thresholds. Furthermore, although this model uses explanatory variables to estimate the location of all thresholds, the way the model is parameterized, without reference to prior thresholds, makes the estimated coefficients more difficult to interpret. Like traditional IRT models, it is important to refer to the prior threshold locations when developing polytomous explanatory models.

1.2.2. *Modeling Polytomous Data*

The notation for the EIRM framework is similar to traditional IRT models. Under the PCM, with adjacent item response categories indexed by *j* and possible item scores from 0 to *J*, the log-odds of selecting response category *j* over *j* – 1 on item *i* $(i = 1, 2, 3, \dots, K)$ for person *n* can be written as:

$$\log\left(\frac{P_n}{P_{n\ (j-1)}}\right) = \theta_n - \left(\delta_\mathrm{i} + \tau_\mathrm{ii}\right), \tag{1}$$

where θ_n represents the latent trait of person *n* and it is normally distributed as $N(\mu_n, \sigma_n^2)$. Traditionally, the $\delta_\mathbf{i}$ is considered an overall index of item difficulty; however, this is actually the location of the threshold between the first (*j* = 0) and second (*j* = 1) response categories for item *i*. The first threshold is often treated as the item difficulty because the first threshold represents the first step to obtain at least a partial credit instead of the lowest possible score on the item. When item response data are dichotomous, there are no estimates for τ_ij. Therefore, a single threshold parameter, $\delta_\mathbf{i}$, becomes the item difficulty parameter. When three or more response categories exist, τ_ij represents the distance between the (*j* – 2)/(*j* – 1) threshold and the (*j* – 1)/*j* threshold for item *i*.

In the explanatory form of PCM (EPCM; Tuerlinckx & Wang, 2004), the log-odds of selecting response *j* over *j* −1 on item *i* for person *n* can be written as:

$$\log\left(\frac{P_n}{P_{n\ (j-1)}}\right) = \mathbf{Z_n}\ \boldsymbol{\theta}_\mathbf{n} - \mathbf{X'_n}\ \boldsymbol{\delta}_\mathbf{i} + \tau_\mathrm{ii}, \tag{2}$$

where $\mathbf{Z}_\mathrm{nij}$ is a matrix that can be used to estimate both fixed- and random-effects related to the person traits. When fitting a traditional IRT model, $\mathbf{Z}_\mathrm{nij}$ would be a vector of ones. For the earlier example with examinees who are either native speakers of English or speak English as a second language, the $\mathbf{Z}_\mathrm{nij}$ matrix could include an additional column of ones (for native speakers) and zeros (for non-native speakers) to estimate a fixed effect for English as a primary language as well as a column of ones to estimate a residual person effect. Similar to $\mathbf{Z}_\mathrm{nij}$, $\mathbf{X}_\mathrm{nij}$ is a matrix of item-related information that describes the characteristics of individual items. With traditional IRT models, a matrix with $K - 1$ columns indicating the item would be used to estimate item difficulties for individual items. In the case of the example above, a vector of ones (for items with images) and zeros (for items without images) can be included as an additional column in $\mathbf{X}_{nij}$ to estimate the impact of the absence or presence of images for person *n* on item *i*. Finally, τ_{ij} in Equation 2 represents the distance between the $(j-2)/(j-1)$ threshold and the $(j-1)/j$ threshold for item *i*, as in the traditional PCM.

Equation 2 illustrates one of the main issues that often occur when utilizing EPCM. While explanatory variables are used to describe the traits of respondents and the difficulty of initial thresholds, the model contains no parameters to describe the common variation beyond the initial threshold parameter. In Equation 2, the other threshold parameters are an afterthought; and thus the model allows for parameters to explain the common variability between initial thresholds, respondents, and their interactions, but fails to do the same for subsequent thresholds. RSM forces thresholds between *J* and *J*−1 to be equidistant for all items, whereas PCM allows for unique estimates of all thresholds across the items. Extending the EIRM framework to all thresholds can allow distances between thresholds to be explained using available covariates.

Natesan, Limbers, and Varni (2010) extended the polytomous EIRM research by applying an explanatory form of GRM to polytomous response data. The study combined the polytomous model with cumulative logits (Tuerlinckx & Wang, 2004) and a 2-level latent regression model (Van den Noortgate & Paek, 2004). The authors compared the fit of two models, a GRM and an explanatory GRM model that was the combination of the 2-level latent regression model and the polytomous model with cumulative logits. This explanatory model contained a person covariate related to emotional quality of life and no additional item covariates. The two models were estimated using both Bayesian likelihood estimation in WINBUGS (Lunn, Thomas, & Spiegelhalter, 2000) and a standard IRT model fitting approach in MULTILOG 7 (Thissen, Chen, & Bock, 2003) on data from a five-item emotional functioning scale. The authors argued that the model with the explanatory person covariate was better because a clearer picture of emotional functioning was obtained when a measure of emotional quality of life was included in the model (Natesan et al., 2010). Although the authors used a person covariate in their explanatory IRT model, their study did not focus on explaining the distance between item thresholds.

The most common application of explanatory IRT models to polytomous response data is the identification of differential item functioning (DIF) in items with three or more response options (Williams & Beretvas, 2006; Vaughn, 2006). These models include an interaction term between a person covariate (e.g., gender) and the initial threshold parameter for a given item. The distance between additional thresholds is estimated with a single threshold parameter when using the explanatory RSM or multiple threshold parameters when using the explanatory PCM (i.e., EPCM). With very little implementation of these models, one of the goals of this study is to display the flexibility of polytomous explanatory IRT models that can help researchers understand the context of the distance between thresholds.

1.2.3. *Alternative Parameterizations*

To recap, there have been very few studies that implemented explanatory IRT models with polytomous response data. There are two major reasons for the scarcity of such studies. First, explanatory IRT models for polytomous response data often require a large number of parameters to be estimated, which may be computationally intensive, especially if the data include many items and respondents. Also, as the number of parameters that the model yields increases, the interpretation of model results becomes more difficult (Bulut et al., 2015). Second, the number of software programs for estimating explanatory models with polytomous data has been limited due to the parameterization of these models. Previous research utilized different software programs for the estimation of explanatory IRT models with polytomous response data, such as WINBUGS (Jiao & Zhang, 2014;

Natesan et al., 2010), HLM (Williams & Beretvas, 2006), and the PROC NLMIXED procedure in SAS (Tuerlinckx & Wang, 2004). However, some of these programs (e.g., HLM and SAS) are only commercially available and the others (e.g., WINBUGS) require a strong understanding of the Bayesian modeling.

To avoid the problems described above, some researchers restructured polytomous response data into dichotomous response data and utilized free software programs that are capable of estimating GLMMs with dichotomous data (e.g., Bulut et al., 2015; De Boeck & Partchev, 2012; Plieninger & Meiser, 2014; Prowker & Camilli, 2007; Scheiblechner, 2009; Verhelst & Verstralen, 2008). However, changing the original structure of data often results in information loss and thus adds additional bias to the inferences made from the estimated models. Alternatively, some researchers maintained the original structure of polytomous data but only focused on explaining the initial threshold – or item location – in the estimation process and ignored subsequent thresholds (e.g., Tuerlinckx & Wang, 2004). However, if explanatory IRT models only examine the initial threshold in the items, potential relationships that may exist across all thresholds could be missed when interpreting the model results.

To solve some of these technical problems, a different parameterization of polytomous IRT models is necessary. Consider the number of thresholds estimated using the traditional RSM and PCM models for a dataset that has *J* response categories for each item $(i = 1, 2, 3, \ldots, K)$. Under the traditional RSM, a total of $J-1$ thresholds need to be estimated beyond the initial thresholds for each item. The estimation of only $J - 1$ parameters is quite restrictive because it assumes that the distance between two thresholds is the same across all items. When researchers assume and thus fit such a model, they have strong a theory about responses. If the researchers fit a polytomous explanatory IRT model with predictors related to their theory, then the model can produce conclusions with greater fidelity.

Under the traditional PCM, a total of $(J - 1) \times K$ threshold parameters will be estimated beyond the initial threshold parameters for each item. The estimation of $(J - 1) \times K$ parameters assumes that the distance between any two thresholds on any item is unique. When researchers fit the PCM, there is no *a priori* theory regarding their models, although the distance between thresholds might be related to item characteristics, respondent characteristics, or an interaction between the two. If a researcher chooses an explanatory IRT model with covariates that explains the distances between thresholds over traditional approaches such as the RSM or PCM, then the researcher is potentially choosing a model that restricts the number of threshold estimates and ties those estimates to an underlying theory. The following section elaborates on these potential models, using a new parameterization.

1.2.4. *Strictly Threshold Explanatory Models*

The Strictly Threshold Explanatory Model (STEM) is a compromise of the RSM and the PCM that utilizes EIRM. Rather than estimating a single distance between thresholds as in the RSM, or the unique distances between item-by-step threshold combinations as in the PCM, the STEM constrains the estimation of the distances between threshold locations based on common item and/or person characteristics. In the STEM model, the initial item thresholds (i.e., item difficulties) are estimated without the use of explanatory variables; however, subsequent distances are estimated using explanatory variables. Consider the earlier example where including graphics on mathematics items and respondents' primary language are likely to affect responses to items and particularly the thresholds. For this example, items are scored in one of three categories *incorrect*, *partially correct*, and *correct*. The STEM can be used to estimate this model as follows:

$$\log\left(\frac{P_n}{P_{n\ (j-1)}}\right) = \theta_n - (\delta_\mathrm{i} + \tau_1(\text{Primary English})_n + \tau_2(\text{Other Language})_n + \tau_3(\text{Images})_i). \qquad (3)$$

In this model, θ_n still represents the trait level of person *n* (i.e., mathematical ability), δ_i represents the initial threshold location (i.e., the *incorrect/partially correct* threshold) for item *i*. The distance between the *incorrect/partially correct* and the *partially correct/correct* threshold is estimated using three parameters, τ_1, τ_2, and, τ_3, which represent the distance between thresholds for primary English speakers controlling for images on items, secondary English speakers controlling for images on items, and the presence of images on items controlling for English language status.

While the STEM is a compromise between the RSM and PCM, the model constraints are beneficial for this particular model. Like the PCM, the STEM does not have fixed step threshold parameters that are equal across all items, rather several step parameters based on features embedded within the items. Like the RSM, the STEM is easier to interpret than the PCM due to fewer numbers of estimated parameters. This interpretation is aided by the use of EIRM and distances between item thresholds are now due to an interaction between the threshold and the item features.

1.2.5. *Explanatory Partial Credit Model*

The STEM utilizes EIRM, but only for restricting, explaining, and measuring thresholds beyond the first threshold and does not use explanatory variables for the location of the initial threshold locations (i.e., the *incorrect/partially correct* thresholds). If the STEM seems appropriate, then using explanatory variables for locating the initial threshold and subsequent thresholds for items is a logical extension. To illustrate the Explanatory Partial Credit Model (EPCM) model, consider the same example for the STEM. For the EPCM, both initial thresholds and distances between thresholds are estimated using explanatory variables. Thus, the EPCM can be written as:

$$\log\left(\frac{P_n}{P_{ni(j-1)}}\right) = \theta_n - (\delta_1(\text{Primary English})_n + \delta_2(\text{Other Language})_n + \delta_3(\text{Images})_i$$
$$+ \tau_1(\text{Primary English})_n + \tau_2(\text{Other Language})_n + \tau_3(\text{Images})_i) \qquad (4)$$

In this model, θ_n, τ_1, τ_2, and, τ_3 have the same meaning as the STEM, and the estimates of δ_1, δ_2, and δ_3 represent the effects of primary English speakers, secondary English speakers, and images in items on the initial threshold locations, respectively. Components present on a particular item and respondent in Equation 4 are additive. Thus, the initial threshold for an item with an image from an examinee whose primary language is English would be located at $\delta_1 + \delta_3$. For the same item, the distance between the *incorrect/partially correct* and *partially correct/correct* thresholds is equal to $\tau_1 + \tau_3$. For this example, a total of seven parameters will be estimated regardless of the number of items on the assessment: three parameters for the initial threshold, three parameters for the distances between the thresholds, and a variance component for respondent trait level. A more generalized formula for the EPCM can be written as:

$$\log\left(\frac{P_n}{P_{n\ (j-1)}}\right) = \mathbf{Z_n}\ \boldsymbol{\theta}_\mathbf{n} - \mathbf{X}'_\mathbf{n}\ \boldsymbol{\delta}_\mathbf{i} + \mathbf{W}'_\mathbf{n}\ \boldsymbol{\tau}_\mathbf{ii}, \qquad (5)$$

where $\mathbf{Z_n}\ \boldsymbol{\theta}_\mathbf{n} - \mathbf{X}'_\mathbf{n}\ \boldsymbol{\delta}_\mathbf{i}$ has the same meaning as the adjacent categories logit model introduced in Equation 2, and $\mathbf{W_n}$ is a matrix of indicator variables used to estimate both fixed- and random-effects related to the distances between thresholds, $\boldsymbol{\tau}_\mathbf{ii}$.

There is an important caveat when fitting explanatory IRT models to polytomous data. In the GLMM framework, the number of item-related parameters that can be estimated for each person group is limited to the total number of item-by-step threshold combinations. For example, if a test has 10 items with four response options, thus three thresholds, then a maximum of 30 parameters (10 items x 3 thresholds) per person group could be estimated for the test (i.e., ten for each step level). Therefore, researchers who utilize explanatory models need to ensure that the explanatory IRT model of interest with item and person covariates is capable of estimating all the parameters given the constraint on the number of parameters to be estimated.

1.2.6. *Cross-Classified Explanatory Partial Credit Model*

The cross-classified EPCM extends the EPCM by including an additional variance component for item difficulty. The result is a model that contains two random-effects, a random-effect for person trait level and a random-effect for item thresholds. Figure 1 displays a network graph describing the cross-classified nature of random items and random persons within the GLMM framework (Beretvas, 2008). By including the random item effect in the cross-classified EPCM, the model acknowledges that additional unaccounted item-related variability exists in the data. The random effect is described as a residual item difficulty because the model already includes item-level predictors. The difficulty of each item can be found by extracting item difficulties from a posterior distribution and combining the values with the relevant item-level predictors. Since these models include fixed-effect predictors for items and an additional random effect that accounts for residual item variability, items can be considered partly random or mixed effects (Van Den Noortgate, De Boeck, & Meulders, 2003).

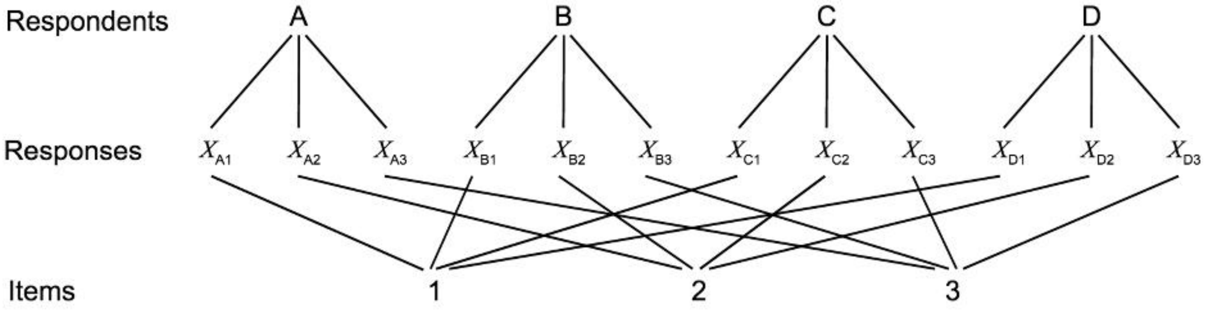

Figure 1. A network graph depicting the cross-classified nature of items and examinees.

As explained earlier, polytomous explanatory IRT models can recover item thresholds either as a fixed effect or as a random effect (Wang, Wilson, & Shih, 2006; Wang & Wu, 2011). From a theoretical standpoint, fitting an IRT model with explanatory item covariates assumes that the researcher has a conceptual understanding of the response process. It is unlikely that the researcher would be able to identify and include all the item-related covariates that can affect the difficulty of an item. The inclusion of the random-effect for item thresholds represents an effect for all of the unexplained components that are not included as fixed explanatory item threshold predictors. Including a random effect suggests that not all features that affect item difficulty are included in the model, but their net effect is a normal distribution of item difficulties with some known mean and variance.

Random item models have been extended to EIRM in several different contexts including, but not limited to, explaining a construct (De Boeck, 2008; Janssen, 2010; Janssen, Schepers, & Peres, 2004), understanding the components of item sets created using automatic item generation (Holling, Bertling, & Zeuch, 2009), predicting item difficulty (Hartig, Frey, Nold, & Klieme, 2012), understanding the impact of cognitive supports on alternative assessments (Ferster, 2013), investigating differential facet functioning (Cawthon, Kaye, Lockhart, & Beretvas, 2012), and modeling item position effects (Albano, 2013). Extending the EPCM in Equation 4, which parameterizes the model for the example considering the role of images in item difficulty for primary and secondary English-speaking students on a mathematics test, the cross-classified EPCM can be written as:

$$\log\left(\frac{P_n}{P_{n\ (j-1)}}\right) = \theta_n - (\delta_1(\text{Primary English})_n + \delta_2(\text{Other Language})_n + \delta_3(\text{Images})_i + \tau_1(\text{Primary English})_n + \tau_2(\text{Other Language})_n + \tau_3(\text{Images})_i + \epsilon_\mathbf{i}) \quad (6)$$

Compared to Equation 4, the only difference in Equation 6 is the additional parameter of $\epsilon_\mathbf{i}$, where $\epsilon_\mathbf{i} \sim N(0, \sigma_i)$, representing the random effect for residual item difficulty. If the estimate of $\epsilon_\mathbf{i}$ is zero, then the model in Equation 6 is equivalent to the EPCM in Equation 4. By including the random item effect in the cross-classified EPCM, the model acknowledges that additional unaccounted item-related variability exists within the data. Since these models include fixed-effect predictors for items and an additional random effect that accounts for residual item variability, items can be considered partly random (Van Den Noortgate et al., 2003).

The explanatory IRT models outlined in this section can be estimated in several ways, most typically marginal maximum likelihood estimation in conjunction with the EM algorithm (Bock & Aitkin, 1981) or restricted maximum likelihood in conjunction with the Laplace estimation. These models can also be estimated using Bayesian methods such as the Markov Chain Monte Carlo estimation method (Gelman, Carlin, Stern, & Rubin, 2013). The aforementioned methods are available through a wide variety of statistical software programs. In this study, we use the *eirm* package (Bulut, 2019) in R (R Core Team, 2018) for estimating traditional and explanatory IRT models for polytomous response data. The *eirm* package is essentially a wrapper for the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015), which is capable of estimating various GLMMs using a restricted maximum likelihood method. In the following sections, we demonstrate how to estimate each of the explanatory IRT models discussed earlier as well as their traditional counterparts using a real dataset with polytomous item responses.

2. METHOD

2.1. Data

For this study, we used a verbal aggression dataset (Vansteelandt, 2000) to demonstrate the estimation of explanatory IRT models with polytomous response data. The verbal aggression dataset consists of the responses to a verbal aggression measure based on potentially frustrating situations. A total of 316 first-year psychology students from a Belgian university (243 women and 73 men) responded to the items about four situations: a bus failing to stop, missing a train, the grocery store closing immediately prior to entering, and an operator disconnecting a call because the respondent can no longer pay. For each situation, the students were asked to decide whether they would curse, shout, or scold; whether they would either do the chosen behavior or just want to do it; and whether they would blame themselves or others for the situation. The situations did not follow a factorial design, but each situation type prompt occurred 6 times resulting in 24 items in total. The response options were *no*, *perhaps*, or *yes* for each item. Table 1 shows a descriptive summary of the items and item covariates in the verbal aggression dataset.

We selected this particular dataset because (1) it is a well-known dataset since it has been used as an illustrative example in previous demonstrations of explanatory IRT models (De Boeck, 2008, 2011; De Boeck & Wilson, 2004, Tuerlinckx & Wang, 2004); (2) it is publicly available through many packages in R (e.g., *lme4* and *difR*) – which would allow readers to replicate the analyses presented in this study (see the Appendix for the R codes); and (3) the small sample size of the verbal aggression dataset justifies the need for estimating a parsimonious model for exploratory purposes rather than a traditional IRT model with many item parameters. Previous research suggested that when polytomous items have three response categories, sample sizes of 300 (or possibly more, if the number of response categories is larger) might be necessary to obtain robust estimates of item threshold parameters (Linacre, 2002; Reise & Yu, 1990). The verbal aggression dataset narrowly exceeds the suggested sample size for polytomous IRT modeling. Therefore, we highlight the trends in the results from traditional IRT models (e.g., RSM and PCM) but intentionally avoid any further interpretation.

Table 1. *Explanatory Variables and Response Frequencies in the Verbal Aggression Dataset*

Item	Situation	Explanatory Variables			Response Options		
		Behavior	Mode	Blame	No	Perhaps	Yes
1	1	Curse	Want	Other	91	95	130
2	1	Scold	Want	Other	126	86	104
3	1	Shout	Want	Other	154	99	63
4	2	Curse	Want	Other	67	112	137
5	2	Scold	Want	Other	118	93	105
6	2	Shout	Want	Other	158	84	74
7	3	Curse	Want	Self	128	120	68
8	3	Scold	Want	Self	198	90	28
9	3	Shout	Want	Self	240	63	13
10	4	Curse	Want	Self	98	127	91
11	4	Scold	Want	Self	179	88	49
12	4	Shout	Want	Self	217	64	35
13	1	Curse	Do	Other	91	108	117
14	1	Scold	Do	Other	136	97	83
15	1	Shout	Do	Other	208	68	40
16	2	Curse	Do	Other	109	97	110
17	2	Scold	Do	Other	162	92	62
18	2	Shout	Do	Other	238	53	25
19	3	Curse	Do	Self	171	108	37
20	3	Scold	Do	Self	239	61	16
21	3	Shout	Do	Self	287	25	4
22	4	Curse	Do	Self	118	117	81
23	4	Scold	Do	Self	181	91	44
24	4	Shout	Do	Self	259	43	14

2.2. Model Overview

The following IRT models were fit the verbal aggression dataset: the RSM, the PCM, the EPCM, and the cross-classified EPCM. All of the models focused on the estimation of the first threshold (i.e., *no*/*perhaps* step) and the second threshold (i.e., *perhaps*/*yes* step) for each item. As explained earlier, the RSM and the PCM are traditional IRT models and thus do not include any item-level or person-level covariates. Note that we included the RSM and PCM for illustrative purposes only; we do not intend to make any inferences from the estimated threshold parameters due to having a small sample size in the verbal aggression dataset. The primary focus of this study was the two explanatory IRT models: the EPCM and the cross-classified EPCM. These models aimed to explain the variability between the step thresholds using item covariates.

Equation 7 shows the RSM and the PCM for the verbal aggression dataset. θ_n represents the overall verbal aggression level of person *n*, δ_i is the initial threshold between the *no* and *perhaps* response categories for item *i*, and τ_ii represents the distance between the *no*/*perhaps* threshold and the *perhaps*/*yes* threshold for item *i*. The only difference between the RSM and the PCM is that τ_ii is the same across all items in the RSM. That is, the distance between the *no*/*perhaps* threshold and the *perhaps*/*yes* threshold is constant across all of the items:

$$\log\left(\frac{P_{n\ (n\ \ )}}{P_{n\ (p\ \ \ ha\ \ \ )}}\right) \text{or} \log\left(\frac{P_{n\ (p\ \ \ ha\ \ \ )}}{P_{n\ (y\ \ )}}\right) = \theta_n - (\delta_\mathrm{i} + \tau_\mathrm{ii}). \tag{7}$$

Equation 8 demonstrates the EPCM with the item-related covariates. In addition to behavior type (i.e., curse, scold, or shout), blame type (others or self) and blame mode (want or do) were used as explanatory covariates in the model. Because the items follow a within-group membership and not between-group membership, all test characteristics cannot be estimated simultaneously because of over-specification. As a result, only a single parameter is needed to estimate the effect of blaming self over blaming others. Similarly, a single parameter is needed to estimate the effect of wanting versus doing an act of verbal aggression.

$$\log\left(\frac{P_{n\,(n\quad)}}{P_{n\,(p\quad ha\quad)}}\right) \text{or} \log\left(\frac{P_{n\,(p\quad ha\quad)}}{P_{n\,(y\quad)}}\right) = \theta_n - (\delta_1(\text{Curse})_{\mathbf{i}} + \delta_2(\text{Scold})_{\mathbf{i}} + \delta_3(\text{Shout})$$

$$+ \delta_4(\text{Do})_{\mathbf{i}} + \delta_5(\text{Self})_{\mathbf{i}} + \tau_1(\text{Curse})_{\mathbf{i}} + \tau_2(\text{Scold})_{\mathbf{i}} + \tau_3(\text{Shout})_{\mathbf{i}}) \qquad (8)$$

The parameters τ_1, τ_2, and, τ_3 in Equation 8 indicate the distances between the *no*/*perhaps* step thresholds and the *perhaps*/*yes* step thresholds. Behavior type (i.e., δ_1, δ_2, and δ_3) also explains the initial step thresholds for the *no*/*perhaps* thresholds. The parameter δ_4 is the difficulty associated with going from wanting to complete a behavior to doing a behavior and δ_5 represents the difficulty associated with going from blaming others to blaming oneself.

Equation 9 demonstrates the cross-classified EPCM with the item-related covariates. This model includes all of the elements in the EPCM in Equation 8. In addition, there is an error term, $\epsilon_{\mathbf{i}}$, which presents the random effect for residual item difficulty.

$$\log\left(\frac{P_{n\,(n\upsilon)}}{P_{n\,(p\quad ha\quad)}}\right) \text{or} \log\left(\frac{P_{n\,(p\quad ha\quad)}}{P_{n\,(y\quad)}}\right) = \theta_n - (\delta_1(\text{Curse})_{\mathbf{i}} + \delta_2(\text{Scold})_{\mathbf{i}} + \delta_3(\text{Shout})_{\mathbf{i}} +$$

$$+ \delta_4(\text{Do})_{\mathbf{i}} + \delta_5(\text{Self})_{\mathbf{i}} + \tau_1(\text{Curse})_{\mathbf{i}} + \tau_2(\text{Scold})_{\mathbf{i}} + \tau_3(\text{Shout})_{\mathbf{i}} + \epsilon_{\mathbf{i}}) \qquad (9)$$

The models summarized in Equations 7 through 9 were fit to the verbal aggression dataset using the *eirm* package (Bulut, 2019). The *eirm* package controls the *glmer* function from the *lme4* package (Bates et al., 2015) and prints model results in a simpler output. The *glmer* function is capable of fitting a GLMM to a dependent variable that follows a binominal distribution within a multilevel structure. Therefore, regular response data in a wide format (persons as rows and items as columns) need to be reformatted into a long format (items nested within persons) and contained indicator codes for items and responses. In addition, polytomous item responses must be transformed into a dichotomous form. The *polyreformat* function from the *eirm* package can transform polytomous items into multiple dichotomous items, without distorting the original response structure. In this study, the response categories of no, perhaps, and yes were dichotomized by creating new labels for each response category. Table 2 shows the reformatted response categories for the verbal aggression dataset.

Table 2. *Reformatting Polytomous Responses into Multiple Dichotomous Responses*

Original Response	Category "perhaps"	Category "yes"
No	0	NA
Perhaps	1	0
Yes	0	1

Because the five IRT models in this study were not nested within each other, a direct comparison between the models using a chi-square test was not possible. Instead, we compared the models using the relative fit indices of the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). The AIC and BIC indices can be calculated using deviance statistics from each model, where deviance is

$$\text{Deviance} = -2(\mathit{loglikelihood}) \tag{10}$$

and AIC and BIC fit indices can be computed as

$$\text{AIC} = \text{Deviance} + (2 \times k)\text{, and} \tag{11}$$

$$\text{BIC} = \text{Deviance} + (df \times \log(n)) \tag{12}$$

where k is the number of estimated parameters and n is sample size. AIC and BIC were chosen for several reasons. First, while AIC and BIC answer two different questions, when the criteria agree on the best model, this provides reassurance on the robustness on the model choice (Kuha, 2004). Second, regardless of the criteria of use for both AIC and BIC, readers have a preferred relative fit index.

3. RESULTS

Table 3 displays the estimated locations of item difficulties and step distances for the RSM and PCM. The RSM is a more restrictive model than the PCM. Item difficulty was estimated for each item individually, while the step distance for *perhaps* to *yes* was fixed (0.54 logits) across the items. The most difficult item based on the location of the *no*/*perhaps* threshold (2.69 logits) was the item S3DoShout, which is about whether the respondent would do a shouting behavior when the grocery store closes just as he or she is about to enter. Also, the RSM indicates that selecting the *yes* option is exp(0.54) = 1.72 times more difficult than selecting *no* and *perhaps* options in the items, after controlling for the latent trait (i.e., verbal aggression) level.

Unlike the RSM, the PCM allows each item to have a unique item difficulty (i.e., the threshold for *no* and *perhaps* options) and a unique step parameter for the distance between the *perhaps* and *yes* option. Based on the estimated item difficulties from the PCM, the item S3DoShout was still the most difficult item (2.71 logits for the *no*/*perhaps* threshold) among the 24 items – which is not surprising given the very low frequency of response option “yes” for this particular item (see Table 1). Unlike the fixed step parameters in the RSM, the PCM had unique step parameters for the distance from *perhaps* to *yes*, ranging from -0.10 to 1.13 across the 24 items. This result suggests that the items in the verbal aggression dataset did not have similar distances from *perhaps* to *yes*, and thus unconstrained step parameters from the PCM can possibly explain more variation among the items.

The PCM results in Table 3 show that four items related to the shouting behavior (S1DoShout, S2DoShout, S4WantShout, and S4DoShout) have a negative distance parameter for perhaps/yes, indicating that the thresholds are not ordered in the same order as the response categories (no, perhaps, and yes). That is, selecting the option “yes” over “perhaps” in these four items was easier for the respondents. This psychometric phenomenon is often called *disordered thresholds* or *reversed deltas* in the literature (Adams, Wu, & Wilson, 2012). In the current study, disordered thresholds may be an indicator of some response processes where respondents prefer to manifest their verbal aggression more explicitly by selecting “yes” rather than “perhaps”. Because the items with disordered thresholds seem to be related to different behavior types (i.e., shouting vs. others), using this characteristic within the EIRM framework can help elucidate the disordered threshold problem.

Table 3. *Locations of No/Perhaps and Perhaps/Yes Thresholds for the Rating Scale Model (RSM) and Partial Credit Model (PCM)*

	RSM				PCM			
Items	Location of no/perhaps		Distance to perhaps/yes		Location of no/perhaps		Distance to perhaps/yes	
	Est.	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*
S1WantCurse	-0.53	0.12	0.54	0.05	-0.38	0.16	0.26	0.21
S1WantScold	-0.13	0.12	0.54	0.05	0.12	0.16	0.00	0.21
S1WantShout	0.35	0.12	0.54	0.05	0.35	0.15	0.56	0.22
S2WantCurse	-0.73	0.12	0.54	0.05	-0.95	0.17	0.89	0.21
S2WantScold	-0.17	0.12	0.54	0.05	0.01	0.16	0.17	0.21
S2WantShout	0.30	0.12	0.54	0.05	0.51	0.15	0.02	0.22
S3WantCurse	0.12	0.12	0.54	0.05	-0.09	0.14	1.02	0.21
S3WantScold	0.96	0.13	0.54	0.05	0.82	0.14	1.02	0.27
S3WantShout	1.54	0.15	0.54	0.05	1.47	0.16	0.90	0.36
S4wantCurse	-0.22	0.12	0.54	0.05	-0.52	0.15	1.13	0.21
S4WantScold	0.65	0.12	0.54	0.05	0.66	0.15	0.49	0.23
S4WantShout	1.12	0.14	0.54	0.05	1.29	0.16	-0.10	0.27
S1DoCurse	-0.42	0.12	0.54	0.05	-0.51	0.16	0.69	0.21
S1DoScold	0.10	0.12	0.54	0.05	0.13	0.15	0.46	0.21
S1DoShout	1.01	0.13	0.54	0.05	1.17	0.16	-0.02	0.26
S2DoCurse	-0.27	0.12	0.54	0.05	-0.15	0.16	0.32	0.21
S2DoScold	0.43	0.12	0.54	0.05	0.45	0.15	0.47	0.22
S2DoShout	1.47	0.14	0.54	0.05	1.63	0.17	-0.09	0.30
S3DoCurse	0.65	0.12	0.54	0.05	0.44	0.14	1.20	0.24
S3DoScold	1.53	0.15	0.54	0.05	1.49	0.16	0.70	0.34
S3DoShout	2.69	0.21	0.54	0.05	2.71	0.22	0.21	0.63
S4DoCurse	0.00	0.12	0.54	0.05	-0.17	0.15	0.91	0.21
S4DoScold	0.69	0.12	0.54	0.05	0.64	0.15	0.66	0.24
S4DoShout	1.88	0.16	0.54	0.05	1.98	0.18	-0.02	0.37

Table 4 displays the estimated item parameters for the EPCM and the cross-classified EPCM. Each model decomposed the *no*/*perhaps* step thresholds based on behavior type (cursing, scolding, shouting), behavior mode (doing or wanting), and blame type (self or others). The main difference between the EPCM and the cross-classified EPCM was the additional random item residuals in the cross-classified EPCM. The top part of Table 4 indicates the location of the *no*/*perhaps* thresholds for the items based on behavior type, behavior mode, and blame type. For instance, the items associated with the cursing behavior type ($\delta_{Curse} = -0.916$) were easier to endorse than the scolding ($\delta_{Scold} = -0.073$) and shouting ($\delta_{Shout} = 0.728$) behavior types for both EPCM and cross-classified EPCM. Also, for the items associated with blaming self over blaming others (δ_{Self}), endorsing the response category of *perhaps* over *no* was exp(0.786) = 2.19 times more difficult in the EPCM and exp(0.82) = 2.27 times more difficult in the cross-classified EPCM. Endorsing *perhaps* over *no* for the mode of doing over the mode of wanting (δ_{Do}) was estimated to be exp(0.465) = 1.59 times more difficult in the EPCM and exp(0.51)=1.67 times more difficult in the cross-classified EPCM.

The bottom part of Table 4 shows the estimated step parameters for the distance from the first threshold (*no*/*perhaps*) to the second threshold (*perhaps*/*yes*), depending on the behavior type. The estimated step parameter for the cursing behavior indicated the largest value for both EPCM (τ_{Curse} = 0.781) and cross-classified EPCM (τ_{Curse} = 0.8). This finding suggests that selecting the response of *yes* over *perhaps* and *no* was more difficult for the items related to cursing than the items related to scolding and shouting. The opposite of this statement is true for the shouting-related items. That is, selecting *yes* over *perhaps* and *no* was easier for the items related to shouting than those related to either cursing or scolding. When the top and bottom parts of Table 4 are compared, the same trend seems to be reversed. The distance from *no/perhaps* to *perhaps/yes* was the smallest for the shouting behavior ($\tau_S = 0.007$) whereas the same distance for the cursing behavior was the largest ($\tau_C = 0.781$). This finding suggests that endorsing *yes* over *perhaps* and *no* in the cursing items required high levels of verbal aggression, whereas endorsing *yes* over *perhaps* and *no* in the shouting items was much easier for the respondents.

Table 4. *Summary of the EPCM and Crossed-Classified EPCM*

	EPCM			Cross-Classified EPCM		
	b	SE	exp(*b*)	*b*	SE	exp(*b*)
Behavior – Curse, Scold or Shout						
Curse	-0.916	0.082	0.400	-0.961	0.134	0.383
Scold	-0.073	0.079	0.930	-0.117	0.132	0.890
Shout	0.728	0.080	2.071	0.714	0.132	2.042
Blame – Self or Others						
Self	0.786	0.047	2.194	0.820	0.105	2.270
Mode – Do or Want						
Do	0.465	0.046	1.592	0.510	0.105	1.665
No/Perhaps to Perhaps/Yes – Step x Behavior						
Step x Curse	0.781	0.076	2.184	0.800	0.077	2.226
Step x Scold	0.395	0.110	1.484	0.440	0.111	1.553
Step x Shout	0.007	0.124	1.007	0.158	0.126	1.171

Table 5 displays the model fit results for the four IRT models. Comparing the RSM to the PCM, AIC favors the PCM, while BIC favors the more parsimonious RSM. Given the disagreement between AIC and BIC, there is not a robust agreement between the relative model fit statistics and thus we cannot make a decision regarding whether the distance between the *no*/*perhaps* and *perhaps*/*yes* should be equidistant across the items. The EPCM used three covariates to explain item difficulties and one covariate to explain step parameters. For the explanatory IRT models, both AIC and BIC favored the cross-classified EPCM, which is not a surprising outcome because the cross-classified EPCM includes more parameters. The model estimates fixed effects for the behavior type, behavior mode, and blaming as well as random effects for the individual items that represent the thresholds of *no* to *perhaps*.

Table 5. *Summary of the Model-Fit Results from the Four IRT Models*

Model	*df*	AIC	BIC
Rating Scale Model	26	11470	11656
Partial Credit Model	49	11450	11801
Explanatory Partial Credit Model	9	11521	11586
Cross-classified Explanatory Partial Credit Model	10	11469	11548

4. DISCUSSION

Traditional polytomous IRT models provide information about the threshold locations of items and estimate the latent trait levels of respondents. Although traditional IRT models are capable of describing respondents and items, they often fail to explain why thresholds for certain items function in a different way. Polytomous explanatory IRT models presented in this study are an alternative that can provide a meaningful context to the response processes.

Previous studies that utilized explanatory IRT models with polytomous data estimated the location of all thresholds without reference to prior thresholds, making the coefficients difficult to interpret. This study took an alternative approach to parameterizing thresholds so that the distance between thresholds would have a simplified interpretation. By improving the interpretation of these parameters, it potentially allows for improved practices in developing measurement instruments – such as surveys, scales, and questionnaires. In addition to re-parameterizing the explanatory IRT models for polytomous data, this study also displayed the versatility of explanatory IRT models by comparing several models.

A total of four traditional and explanatory IRT models were fit to the verbal aggression data: RSM, PCM, EPCM, and cross-classified EPCM. AIC and BIC model fit indices were used to compare the models. Both model-fit indices favored explanatory IRT models over the more traditional RSM and PCM. The AIC statistic favored the cross-classified EPCM, which included both the first threshold parameters as random effects and additional covariates to explain the distance between the no/perhaps and perhaps/yes step threshold locations. When comparing the relative fit of the cross-classified EPCM to the other models, the findings suggested that cross-classified models could be useful, as the model showed better relative fit compared to its more restrictive EPCM counterpart. Despite the inconclusive relative fit of the cross-classified EPCM and EPCM, these models show great utility in explaining why an item may be difficult by using information from the collective assessment. For instance, the explanatory item response models indicated that respondents are less likely to select a *yes* response for an item associated with the shouting behavior than other behavior types, which is the type of information that would not otherwise be collected from either the RSM or the PCM.

While EIRM has not been typically used for explaining the distance between step thresholds in polytomous items, this study revealed a situation where the estimated step thresholds between response categories did not vary enough for some items. For instance, after fitting explanatory models to explain the difference between the location of the *no*/*perhaps* response threshold and the *perhaps*/*yes* threshold, the results showed that a group of items, specifically the items sharing the same shout behavior type did not show a statistically significant difference between the no/perhaps and the *perhaps*/*yes* step thresholds. This could be interpreted as the *perhaps* option not being as useful to understanding respondents underlying verbal aggression for the items where verbal aggression is expressed through shouting. This means respondents were likely to skip perhaps and go from selecting no to yes when reaching a certain level of aggression. This finding implies that when the more traditional PCM is fit to the data, there are instances where *perhaps* thresholds do not function properly. By using explanatory response models, the functionality of multiple response categories in polytomous items can be determined and the cases where of a response option not functioning could be explained by using an item-related covariate.

While this study only examined polytomous data structures with explanatory item response models, future studies can compare the conclusions drawn from polytomous explanatory IRT models against the findings from the models where polytomous item responses are dichotomized and dichotomous IRT models are applied. Additionally, the models in this study can be modified for different purposes, such detecting item parameter drift and construct shift in polytomously-scored items and improving test equating/linking results in both dichotomous and polytomous data settings.

ORCID

Luke Stanke https://orcid.org/0000-0001-5853-1267
Okan Bulut https://orcid.org/0000-0002-4340-6954

5. REFERENCES

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*(4), 408–426. doi:10.1111/jedm.12026

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547–573. doi: 10.1177/0013164411432166

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4) 581–594. doi:10.1177/014662167800200413

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6)*, 716–723. doi:10.1109/TAC.1974.1100705

Bates, D., Maechler, M., Bokler, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01

Beretvas, S. N. (2008). Cross-classified random effects models. In A. A. O'Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 161-197). Charlotte, SC: Information Age Publishing.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. doi:10.1007/BF02291411

Bock, R. D., & Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. doi:10.1007/BF02293801

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

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89 - 118. http://dx.doi.org/10.1080/08957340801926086

Bulut, O. (2019). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses* [Computer software]. Available from https://github.com/okanbulut/eirm.

Bulut, O., Palma, J., Rodriguez, M. C., & Stanke, L. (2015). Evaluating measurement invariance in the measurement of developmental assets in Latino English language groups across developmental stages. *Sage Open, 5*(2), 1-18. doi:10.1177/2158244015586238

Cawthon, S., Kaye, A., Lockhart, L., & Beretvas, S. N. (2012). Effects of linguistic complexity and accommodations on estimates of ability for students with learning disabilities. *Journal of School Psychology, 50*, 293–316. doi:10.1016/j.jsp.2012.01.002

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133–148. doi:10.1111/j.1745-3984.2005.00007

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*(3-4), 243–276. http://dx.doi.org/10.1080/15305058.2002.9669495

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533-559. doi:10.1007/s11336-008-9092-x

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28.

De Boeck, P., & Wilson, M. (2004). Explanatory item response models: a generalized linear and nonlinear approach. *Statistics for Social Science and Public Policy*. New York, NY. Springer.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Boca Raton, FL: CRC Press.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1). 179–197. http://dx.doi.org/10.1037/0033-2909.93.1.179

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds, *Cognitive Assessment* (pp. 107–135). Springer USA.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396. http://dx.doi.org/10.1037/1082-989X.3.3.380

Embretson, S. E. (2006*). Cognitive models for the psychometric properties of GRE quantitative items*. Final Report. Princeton, NJ: Educational Testing Service.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Yang, X. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 119–145). New York, NY: Cambridge University Press.

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

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*(3). 299–317. doi:10.1111/j.1745-3984.2010.00115.x

Ferster, A. E. (2013). *An evaluation of item level cognitive supports via a random-effects extension of the linear logistic test model*. Unpublished doctoral dissertation, University of Georgia.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis.* Boca Raton, FL: CRC Press.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*(4), 665–686. doi:10.1177/0013164411430707

Holling, H., Bertling, J. P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, *35*, 71–76. doi:10.1016/j.stueduc.2009.10.004

Janssen, R. (2010). Modeling the effect of item designs within the Rasch model. In. S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 227–245). Washington, DC, US: American Psychological Association.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York, NY: Springer-Verlag.

Jiao, H., & Zhang, Y. (2014). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology*, *68*(1), 65–83. doi:10.1111/bmsp.12035

Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing*, *14*(3), 245–264. http://dx.doi.org/10.1080/15305058.2013.877911

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions of performance. *Sociological Methods and Research*, *33*, 188–229. doi:10.1177/0049124103262065

Kubinger, K. (2008). On the revival of the Rasch model-based LLTM: from constructing tests using item generating rules to measuring item administration effects. *Psychological Science Quarterly, (3)*, 311–327.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *5*(1), 85–106.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337. doi:10.1023/A:1008929526011

Luppescu, S. (2012, April). *DIF detection in HLM item analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Natesan, P., Limbers, C., & Varni, J. W. (2010). Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. *Educational and Psychological Measurement*, *70*(3) 420–439. doi:10.1177/0013164409355696

Plieninger, H. & Meiser, T. (2014). Validity of multi-process IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. doi:10.1177/0013164413514998

Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journal of Educational Measurement*, *44*(1), 69–87. doi:10.1111/j.1745-3984.2007.00027.x

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. doi:10.1214/aos/1176344136

Scheiblechner, H. H. (2009). Rasch and pseudo-Rasch models: suitableness for practical test applications. *Psychology Science Quarterly*, *51*, 181–194.

Thissen, D., Chen, W., & Bock, D. (2003). *MULTILOG 7* [Computer software]. Chicago, IL: Scientific Software International.

Tuerlinckx, F., & Wang, W.-C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75–109). New York: Springer-Verlag.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55.

Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics and Data Analysis*, *11*(3), 275–295. doi:10.1111/j.2044-8317.1990.tb00925.x

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach*. Electronic Theses, Treatises and Dissertations. Paper 4588.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*(4), 369–386. doi:10.3102/10769986028004369

Van den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 167–187). New York, NY: Springer-Verlag.

van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). New York: Springer

Vansteelandt, K. (2000). *Formal models for contextualized personality psychology*. Unpublished doctoral dissertation, K.U. Leuven, Belgium.

Verhelst, N. D., & Verstralen, H. H. F. M. (2008). Some considerations on the Partial Credit Model. *Psicologica: International Journal of Methodology and Experimental Psychology*, *29*(2), 229–254.

Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement, 67*(4), 583 - 605. doi:10.1177/0013164406296974

Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296 - 318. doi:10.1177/0146621605276281

Wang, W.-C., Wilson, M., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement*, *43*(4), 335–353. doi:10.1111/j.1745-3984.2006.00020.x

Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement*, *48*(4), 441-456. doi:10.1111/j.1745-3984.2011.00154.x

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement*, *30*, 22–42. doi:10.1177/0146621605279867

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In Hartig, J., Klieme, E., Leutner, D. (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects* (pp. 91-120). Göttingen, Germany: Hogrefe & Huber.

Wilson, M., Zheng, X., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, *13*(1), 1–22.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*(4), 589–600.

Appendix

R codes for estimating the explanatory IRT models with the verbal aggression dataset

```
# Install and load the required packages
install.packages("devtools")
devtools::install_github(repo = "okanbulut/eirm")
library("eirm")


# Reformat the VerbAgg dataset for polytomous EIRM
data("VerbAgg")
VerbAgg2 <- polyreformat(data=VerbAgg, id.var = "id", long.format = FALSE,
                    var.name = "item", val.name = "resp")

# Rating Scale Model
mod1 <- eirm(formula = "polyresponse ~ -1 + item + polycategory + (1|id)",
          data = VerbAgg2)

print(mod1, Easiness = FALSE)


# Partial Credit Model
mod2 <- eirm(formula = "polyresponse ~ -1 + item + item:polycategory +
          (1|id)", data = VerbAgg2)

print(mod2, Easiness = FALSE)


# Explanatory Partial Credit Model
mod3 <- eirm(formula = "polyresponse ~ -1 + btype + situ + mode +
          polycategory + polycategory:btype + (1|id)",
          data = VerbAgg2)

print(mod3, Easiness = FALSE)


# Cross-Classified Explanatory Partial Credit Model
mod4 <- eirm(formula = "polyresponse ~ -1 + btype + situ + mode +
              polycategory + polycategory:btype + (1|item) + (1|id)",
          data = VerbAgg2)

print(mod4, Easiness = FALSE)
```