Article

Dependent Dirichlet Process Rating Model

Applied Psychological Measurement
2014, Vol. 38(3) 217–228

DOI: 10.1177/0146621613512018
apm.sagepub.com
SAGE

Ken Akira Fujimoto[1] and George Karabatsos[1]

Abstract

Typical item response theory (IRT) rating scale models assume that the rating category threshold parameters are the same over examinees. Unfortunately, such models are inappropriate for rating data that exhibit differential item functioning (DIF). The authors introduce a new Bayesian nonparametric IRT model for rating scale items, which is more appropriate for rating data that contain DIF. The model is an infinite mixture of Rasch partial credit models, with mixture distribution modeled by the local (Dependent) Dirichlet process. The model treats the rating category thresholds as the random parameters that are subject to the mixture, with (stick-breaking) mixture weights that are covariate-dependent. Thus, the model allows the rating category thresholds to differ across items and examinees, while allowing the form of the distribution for the category thresholds to vary flexibly as a function of covariates. The authors illustrate the new model through the analysis of simulated data and real data. The model demonstrated the ability to correctly identify DIF items. Moreover, the model attained better predictive-fit performance than did other commonly used IRT rating models.

Keywords

rating scale analysis, Bayesian nonparametrics, Bayesian inference, differential category usage, differential item functioning

In social science research, it is often of interest to analyze examinee ratings to items of a test. Item response theory (IRT) rating models are commonly used for this purpose because of the useful information they provide about the psychometric properties of the rating data. For example, they provide information about the overall difficulty and category threshold parameters of each test item, and the test ability (latent trait) parameter of each examinee. Typical IRT rating models include the Rasch Rating Scale Model (RSM; Andrich, 1978), partial credit models (PCM; Masters, 1982; Muraki, 1992), and the graded response model (GRM; Samejima, 1969, 1972), all of which have seen many successful applications in a wide range of research settings.

Nevertheless, these IRT models have their limitations. Typical IRT rating models assume that the same rating category threshold parameters apply to all examinees. However, this assumption is violated when differential rating category usage occurs across the examinees.

[1]University of Illinois at Chicago, USA

Corresponding Author:
Ken Akira Fujimoto, Department of Educational Psychology, University of Illinois at Chicago, 3343 EPASW, 1040 W. Harrison, Chicago, IL 60607, USA.
Email: kfujim3@uic.edu

Differential rating category usage may be caused by differential item functioning (DIF), that is, when different clusters (groups) of examinees give rise to different threshold estimates for the rating categories after controlling for the level of examinee ability. The different clusters could either refer to unknown latent groups or known examinee groups (e.g., male and female). Differential category usage across examinees could also arise from nonsystematic random error, such as when unclear labels that are assigned to the rating categories. Regardless, if a typical IRT model is used to analyze data that violate its assumption of no differential category usage, then the model may poorly fit the data and produce misleading results. The results would wrongly indicate that, for each test item, a single set of rating category threshold estimates applies for all examinees. In turn, this could lead to misleading examinee ability estimates. With traditional models, item fit statistics are often relied upon to identify items that misfit the model. However, fit statistics have low power in identifying DIF items (Seol, 1999; Smith & Suh, 2003). Moreover, even when an item fit statistic identifies an item as problematic, it does not explain why the item is misfitting.

Multiple-group IRT models (e.g., Lord, 1980; Wright & Masters, 1982) are more appropriate when the differential category usage is a result of DIF. These models specify interaction covariates between person and item characteristics (e.g., overall item difficulty and category thresholds). The regression coefficients of these interaction terms indicate whether DIF is present in an item and provide some explanation about how rating category usage varies as a function of examinee characteristics. This modeling approach, however, is still limited because it assumes that the model contains all the covariates that could explain DIF. As mentioned, latent or unknown examinee characteristics may also contribute to differential rating category usage, and/or random error may be present in the rating thresholds.

It then seems preferable to specify a discrete mixture IRT rating model that can identify and account for differential rating category usage in the items, which may either result from multiple latent clusters (groups) of examinees, and/or result from known examinee characteristics (covariates). For each item, and conditioned on any other known covariates, the model would specify a (mixture) distribution for the rating category thresholds over all examinees, while assigning a distinct set of rating category threshold parameters to each latent cluster of examinees. If all examinees use (e.g., interpret) an item's rating categories in the same manner, then the model's threshold distribution becomes unimodal with near-zero variance. Such a distribution would indicate a single cluster of examinees in terms of the rating thresholds, as in typical IRT rating models which assume no differential category usage. When an item exhibits differential rating category usage over examinees, then the model's threshold distribution will have noticeable variance, with possible skewness and/or multimodality. A unimodal distribution with noticeable variance and/or skewness could either indicate uncertainty in the rating category usage of the item or DIF. A multimodal distribution would indicate DIF and multiple latent clusters of examinees. Finally, when an item's threshold distribution is shown to depend on one or more known covariates that describe examinee background characteristics (e.g., gender, income), after controlling for examinee ability, then there is DIF due to known examinee groupings (as in multiple-group IRT).

A discrete mixture model has the general form (e.g., McLachlan & Peel, 2000):

$$f_{G_{\mathbf{x}}}(y|\mathbf{x}) = \int f(y|\mathbf{x};\boldsymbol{\xi},\boldsymbol{\Psi}(\mathbf{x}))\mathrm{d}G_{\mathbf{x}}(\boldsymbol{\Psi}) = \sum_{h=1}^{H} f(y|\mathbf{x};\boldsymbol{\xi},\boldsymbol{\Psi}_h(\mathbf{x}))\omega_h(\mathbf{x}),$$

given a mixing distribution $G_{\mathbf{x}}$ that is possibly covariate ($\mathbf{x}$) dependent; component indices $h = 1, \ldots, H$, kernel (component) densities $f(y|\mathbf{x};\boldsymbol{\xi},\boldsymbol{\Psi}_h(\mathbf{x}))$ ($h = 1, \ldots, H$) with fixed parameters $\boldsymbol{\xi}$ and random parameters $\boldsymbol{\Psi}_h(\mathbf{x})$ that are subject to the mixture; and given mixing weights

$(\omega_h(\mathbf{x}))_{h=1}^{H}$ which sum to 1 at every $\mathbf{x} \in \mathcal{X}$. Mixture IRT models treat $y \in \{k = 0, 1, \ldots, m\}$ as a scored item response (e.g., a rating), and specify each of the kernel densities $f(y|\mathbf{x}; \boldsymbol{\xi}, \boldsymbol{\Psi}_h(\mathbf{x}))$ by an ordinary IRT model, such as a two-parameter logistic model, or a RSM.

Typical IRT mixture models assume finite mixtures (i.e., $H < \infty$; Frick, Strobl, Leisch, & Zeileis, 2012; Rost, 1991; Smit, Kelderman, & van der Flier, 2003; von Davier & Yamamoto, 2004), which limits their ability to adequately describe many rating scale data sets. The authors could achieve greater modeling flexibility by turning to a fully nonparametric framework, through the specification of an infinite-mixture model (i.e., $H = \infty$). Such a model has infinitely-many parameters, thus avoiding the restrictive assumption of parametric IRT models, namely, that the distribution of item response data can be fully described by finitely-many parameters. Along these lines, infinite-mixture IRT models have been developed. They include models based on the Dirichlet process (DP) mixture of the item parameters of a three-parameter logistic model (Miyazaki & Hoshino, 2009), models based on a DP mixture of ability parameters in a Rasch model (San Martín, Jara, Rolin, & Mouchart, 2011), and a Dependent Dirichlet process (DDP) mixture model for the link function of the two-parameter IRT model (Duncan & MacEachern, 2008). Karabatsos and Walker (in press) review the DP and DDP mixture models for IRT. However, none of the available mixture IRT models provide clustering of examinees in terms of the rating category threshold parameters. This is because they do not treat the rating category threshold parameters as the random parameters (i.e., the $\boldsymbol{\Psi}_h(\mathbf{x})$) that are subject to the mixture.

To address the limitations of the existing IRT models, the authors introduce a novel Bayesian nonparametric IRT rating model, which they call the DDP Rating Model (DDP-RM). This model is an infinite mixture of Rasch PCM, with rating category threshold parameters subject to the mixture, and with covariate-dependent stick-breaking weights. The random parameters and the mixture weights are modeled by a novel version of the local Dirichlet process (lDP; Chung & Dunson, 2011), which is a DDP (MacEachern, 1999, 2000, 2001).

In ''The DDP-RM'' section, the authors introduce the DDP-RM. In the ''Illustration of the DDP-RM on Simulated Data'' section, they illustrate their model on simulated data, to demonstrate the model's ability to identify DIF as a result of latent (unknown) examinee characteristics (covariates). In the ''Illustration of the DDP-RM on Real Data'' section, the authors illustrate their model on a real data set of rating scale items, which has been extensively studied in the psychometric modeling literature (De Boeck & Wilson, 2004). In this illustration, the authors also compare the goodness of predictive fit between the DDP-RM and other IRT rating scale model that are commonly utilized. In the ''Conclusion'' section, they conclude by discussing possible future extensions of their model. Throughout, they denote $\mathrm{n}(\cdot|\cdot,\cdot)$, $\mathrm{n}_p(\cdot|\cdot,\cdot)$, $\mathrm{ga}(\cdot|\cdot,\cdot)$, $\mathrm{ig}(\cdot|\cdot,\cdot)$, $\mathrm{beta}(\cdot|\cdot,\cdot)$, and $\mathrm{un}(\cdot|\cdot,\cdot)$ as the probability density functions for the univariate normal, p-variate Normal, gamma, inverse gamma, beta, and uniform distributions, respectively. The gamma and inverse gamma distributions are parameterized by shape and rate parameters.

The DDP-RM

The rating model, the DDP-RM, is defined by an infinite mixture of IRT rating model. Specifically, this mixture model assumes that the probability of a rating is defined by

$$P(Y = y|\mathbf{x}; \theta, \boldsymbol{\upsilon}, \boldsymbol{\gamma}, \boldsymbol{\psi}) = \int f(y|\theta, \boldsymbol{\tau}) \mathrm{d}G_{\mathbf{x}}(\boldsymbol{\tau}) = \sum_{h=1}^{\infty} f(y|\theta_t, \boldsymbol{\tau}_h) \omega_h\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}; \boldsymbol{\upsilon}, \boldsymbol{\gamma}, \boldsymbol{\psi}\right), \quad (1)$$

with kernel probability densities $f(y|\theta, \boldsymbol{\tau}_h)$ specified by the PCM,

$$f(y|\theta,\boldsymbol{\tau}_h)=P(Y=y|\theta,\boldsymbol{\tau}_h)=\frac{\exp\left(y\theta-\sum_{l=0}^{y}\boldsymbol{\tau}_{lh}\right)}{\sum_{k=0}^{m}\exp(k\theta-\sum_{l=0}^{k}\boldsymbol{\tau}_{lh})},\ h=1,2,\ldots, \tag{2}$$

where the mixture distribution $G_{\mathbf{x}}$ is covariate ($\mathbf{x}$) dependent and defined by

$$G_{\mathbf{x}}(\cdot)=\sum_{h=1}^{\infty}\omega_h\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}\right)\delta_{\boldsymbol{\tau}_h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})}(\cdot) \tag{3}$$

and where $\delta_{\boldsymbol{\tau}}(\cdot)$ denotes the degenerate distribution which assigns probability 1 to the value $\boldsymbol{\tau}$. In addition, θ_t denotes the ability parameter of a given examinee t, for a sample of examinees indexed by $t=1,2,\ldots,N$; and for the $m+1$ rating categories indexed by $k=0,1,\ldots,m$, the vector $\boldsymbol{\tau}_h=(\tau_{1h},\ldots,\tau_{mh})^{\mathrm{T}}$ gives the set of rating category threshold parameters for the hth mixture component, while assuming the constraint $\tau_{0h}\equiv 0$. The mixture distribution $G_{\mathbf{x}}(\boldsymbol{\tau})$ for the thresholds, and the corresponding covariate ($\mathbf{x}$)-dependent mixture weights $\{\omega_h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})\}_{h=1,2,\ldots}$ and atoms $\{\boldsymbol{\tau}_h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})\}_{h=1,2,\ldots}$, are modeled by a modified lDP prior. Therefore, the mixture weights have a stick-breaking form (see Sethuraman, 1994); later, the authors provide more details about the lDP and these weights. In general, $\mathbf{x}$ can be a vector of any p covariates, $\mathbf{x}=(x_1,\ldots,x_p)$, and they, respectively, correspond to (positive-valued) linear regression coefficients $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_p)^{\mathrm{T}}$. For example, the covariates could be dummy (0-1) test item indicators, describe examinee characteristics (e.g., gender, race, and/or social economic status), and/or describe other test characteristics (e.g., time at which item was administered, item type, etc.).

As shown in Equation 3, the DDP-RM is based on an infinite-mixture distribution $G_{\mathbf{x}}(\boldsymbol{\tau})$ for the rating category thresholds $\boldsymbol{\tau}$. Therefore, conditionally on $\mathbf{x}$, the model can account for virtually all distributions of the rating category thresholds ($\boldsymbol{\tau}_h$). These distributions include unimodal distributions with small variance, indicating an item is free of DIF; unimodal distributions with larger variance and/or skewness, indicating an item with more uncertainty in rating category usage, and possibly DIF; and multimodal distributions, which indicate the presence of multiple latent clusters of examinees (i.e., DIF). In addition, the shape and location of the mixture distribution $G_{\mathbf{x}}$ can change flexibly as a function of the covariates ($\mathbf{x}$). Therefore, at one extreme, the mixture distribution $G_{\mathbf{x}}$ may be unimodal with small variance for one value of the covariates $\mathbf{x}$, while for the other extreme, the mixture distribution $G_{\mathbf{x}'}$ may be highly skewed and multimodal for a different value of the covariates $\mathbf{x}'$.

The mixture distribution, $G_{\mathbf{x}}$, of the model is formed according to the authors' following novel modification of the lDP (Chung & Dunson, 2011), which is described as follows. First let

$$\mathcal{L}_{\mathbf{x}}=\left\{h:d\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma},h\right)\le\psi(\mathbf{x})\right\}\subset\{1,2,\ldots\}$$

be the subset of mixture component indices $h\in\mathbb{Z}+$ having fixed addresses $\{\Gamma_h\equiv h\}$ that are within a $\psi(\mathbf{x})$-neighborhood around the linear predictor $\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}$, $\pi_l(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})$ is the lth ordered index in $\mathcal{L}_{\mathbf{x}}$, and $d(\cdot,\cdot)$ is a chosen distance measure (e.g., Euclidean). For example, if $\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}=10$ and $\psi(\mathbf{x})=2.5$, then the covariate ($\mathbf{x}$)-dependent local subset becomes $\mathcal{L}_{\mathbf{x}}=\{8,9,10,11,12\}$, and $\pi_1(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})=8,\pi_2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})=9,\ldots,\pi_{|\mathcal{L}_{\mathbf{x}}|}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})=12$, where $|\mathcal{L}_{\mathbf{x}}|$ is the cardinality of the set $\mathcal{L}_{\mathbf{x}}$. Under the formulation of the lDP, the local variables are defined by $\boldsymbol{\upsilon}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})=\{\upsilon_h,h\in\mathcal{L}_{\mathbf{x}}\}$, to specify the mixture weights in Equation 3 as having the covariate-dependent, stick-breaking form:

$$\omega_l\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}\right)=\upsilon_{\pi_l(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})}\prod_{r<l}\left(1-\upsilon_{\pi_r(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})}\right), \tag{4}$$

where the rating threshold atoms $\boldsymbol{\tau}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})=\{\boldsymbol{\tau}_h,h\in\mathcal{L}_{\mathbf{x}}\}$ are also covariate-dependent. The authors fix $\upsilon_{\max(\mathcal{L}_{\mathbf{x}})}(\mathbf{x}'\boldsymbol{\gamma})\equiv 1$ to ensure that the mixture weights $\omega_l(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})$ sum to 1 for each $\mathbf{x}$

(Chung & Dunson, 2011). In short, the lDP forms stick-breaking mixture weights by selecting the strict subset of stick-breaking parameters ($\{\upsilon_h\}$) and atoms ($\{\boldsymbol{\tau}_h\}$) that are within the neighborhood centered around (a linearized) $\mathbf{x}$. Then the mixture weights of Equation 4 gives rise to a covariate-dependent mixing distribution in Equation 3, which can be rewritten as follows:

$$G_{\mathbf{x}}(\cdot) = G_{\mathbf{x}}(\cdot; \boldsymbol{\tau}, \boldsymbol{\upsilon}, \boldsymbol{\gamma}, \boldsymbol{\psi}) = \sum_{l=1}^{|\mathcal{L}_{\mathbf{x}}|} \omega_l\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}\right)\delta_{\boldsymbol{\tau}_{\pi_l(\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma})}}(\cdot), \tag{5}$$

where the authors denote $\boldsymbol{\tau} = (\boldsymbol{\tau}_h)_{h=1}^{\infty}$, $\boldsymbol{\upsilon} = (\upsilon_h)_{h=1}^{\infty}$, and $\boldsymbol{\psi} = (\psi(\mathbf{x}))_{\mathbf{x}\in\mathcal{X}}$. Based on this specification, for two covariates, $\mathbf{x}$ and $\mathbf{x}'$, the level of similarity between $\mathcal{L}_{\mathbf{x}}$ and $\mathcal{L}_{\mathbf{x}'}$ determines the level of similarity between the two corresponding mixing distribution, $G_{\mathbf{x}}(\cdot)$ and $G_{\mathbf{x}'}(\cdot)$, with the level of similarity controlled by the parameters $(\boldsymbol{\gamma}, \boldsymbol{\psi})$.

The DDP-RM is completed by the specification of the following prior distributions:

$$\begin{aligned}
\theta_t &\sim \mathrm{n}(0, \sigma^2), t = 1, 2, \ldots, N;\\
\sigma^2 &\sim \mathrm{ig}(\sigma^2 | a_{\sigma^2}, b_{\sigma^2});\\
\boldsymbol{\tau}_h, \upsilon_h &\sim \mathrm{n}_{m_j}(\boldsymbol{\tau}|\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\tau}})\mathrm{beta}(\upsilon|1, \alpha), h = 1, 2, \ldots;\\
\boldsymbol{\alpha}, \boldsymbol{\gamma} &\sim \mathrm{ga}(\alpha|a_{\alpha}, b_{\alpha}) \textstyle\prod_{j=1}^{p} \mathrm{un}\left(\gamma_j | a_{\gamma}, b_{\gamma}\right);\\
\psi(\mathbf{x}) &\sim \mathrm{un}\left(a_{\psi}, b_{\psi}\right), \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

If so desired, one may fix various model parameters to a particular constant by making specific extreme choices of prior. For example, we can fix $\psi(\mathbf{x})$ to a constant c by setting $a_{\psi} = b_{\psi} = c$ in the uniform prior.

Bayesian Posterior Inference of the DDP-RM

For notational convenience, the authors denote a sample set of rating data by $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n=NJ}$, provided by N examinees ($t = 1, \ldots, N$) on J test items ($j = 1, \ldots, J$), and with $n = NJ$ giving the total number of item responses in the data set. Each y_i denotes a rating by a particular examinee on a particular item. In addition, as before, they denote the parameters of their model by $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \sigma^2, \boldsymbol{\tau}, \boldsymbol{\upsilon}, \alpha, \boldsymbol{\gamma}, \boldsymbol{\psi})$, with $\boldsymbol{\theta} = (\theta_t)_{t=1}^{N}$, $\boldsymbol{\tau} = (\boldsymbol{\tau}_h)_{h=1}^{\infty}$, $\boldsymbol{\upsilon} = (\upsilon_h)_{h=1}^{\infty}$, $\boldsymbol{\gamma} = (\gamma_k)_{k=1}^{p}$, and $\boldsymbol{\psi} = (\psi(\mathbf{x}))_{\mathbf{x}\in\mathcal{X}}$.

According to Bayes's theorem, given a data set $\mathcal{D}_n$ and having likelihood $\prod_{i=1}^{n} f(y_i|\mathbf{x}_i; \boldsymbol{\zeta})$ under the model with parameters $\boldsymbol{\zeta}$, with a proper prior density $\pi(\boldsymbol{\zeta})$ defined over the space $\boldsymbol{\Omega}_{\boldsymbol{\zeta}}$ of $\boldsymbol{\zeta}$, the posterior density of $\boldsymbol{\zeta}$ is proper and is given by

$$\pi(\boldsymbol{\zeta}|\mathcal{D}_n) \propto \textstyle\prod\prod_{i=1}^{n} P(y_i|\mathbf{x}_i; \boldsymbol{\zeta})\pi(\boldsymbol{\zeta})$$

up to a proportionality constant. Then the posterior predictive density of Y for a chosen $\mathbf{x}$ is given by

$$f_n(y|\mathbf{x}) = \int f(y|\mathbf{x}; \boldsymbol{\zeta})\pi(\boldsymbol{\zeta}|\mathcal{D}_n)\mathrm{d}\boldsymbol{\zeta},$$

and this density corresponds to posterior predictive mean (expectation) $\mathrm{E}_n(Y|\mathbf{x}) = \int y f_n(y|\mathbf{x})\mathrm{d}y$ and variance (Var) $\mathrm{Var}_n(Y|\mathbf{x}) = \int \{y - \mathrm{E}(Y|\mathbf{x})\}^2 f_n(y|\mathbf{x})\mathrm{d}y$. In addition, when investigating for DIF, it is of interest to infer functionals of the posterior predictive mean $\mathrm{E}_n[G_{\mathbf{x}}(\cdot)]$ of the threshold mixture distribution $G_{\mathbf{x}}(\boldsymbol{\tau})$, such as its density. This posterior predictive mean is defined by

$$\mathrm{E}_n[G_{\mathbf{x}}(\cdot)] = \int\int\int\int G_{\mathbf{x}}(\cdot;\boldsymbol{\tau},\boldsymbol{\upsilon},\boldsymbol{\gamma},\boldsymbol{\psi})\pi(\boldsymbol{\tau},\boldsymbol{\upsilon},\boldsymbol{\gamma},\boldsymbol{\psi}|\mathcal{D}_n)\mathrm{d}\boldsymbol{\tau}\mathrm{d}\boldsymbol{\upsilon}\mathrm{d}\boldsymbol{\gamma}\mathrm{d}\boldsymbol{\psi},$$

given the marginal posterior density $\pi(\boldsymbol{\tau},\boldsymbol{\upsilon},\boldsymbol{\gamma},\boldsymbol{\psi}|\mathcal{D}_n) = \int\int\int \pi(\boldsymbol{\zeta}|\mathcal{D}_n)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\sigma^2\mathrm{d}\alpha$. To perform inference of functionals of the posterior density $\pi(\boldsymbol{\zeta}|\mathcal{D}_n)$, including marginal posterior densities, posterior predictive densities $f_n(y|\mathbf{x})$, the posterior mean mixing distribution $\mathrm{E}_n[G_{\mathbf{x}}(\cdot)]$, the authors make use of standard Markov chain Monte Carlo (MCMC) sampling methods for Bayesian infinite-mixture models (Kalli, Griffin, & Walker, 2011). Online Appendix A provides more details.

Unique Features of the DDP-RM

A unique feature of the DDP-RM is that it allows the mixing distribution, G_x, to change flexibly as a function of the covariates $\mathbf{x}$. In addition, conditionally on any $\mathbf{x}$, the mixture distribution $G_{\mathbf{x}}$ can take on any shape, ranging from unimodal with small variance to highly multimodal with large variance. This flexibility is enabled by a nonparametric specification of the mixing distribution $G_{\mathbf{x}}$ according to a flexible infinite mixture (involving an infinite number of parameters), with covariate-dependent mixture weights (i.e., the ω_h) and thresholds (i.e., the $\boldsymbol{\tau}_h$), as shown in Equations 3 and 5. In other words, the model does not make any parametric assumptions about $G_{\mathbf{x}}$ that traditional models do, such as assume that the mixing distribution is normally distributed. Such parametric assumption implies the empirically falsifiable assumption that the mixing distribution is symmetric and unimodal. The DDP-RM, which is free from such limited assumptions about the mixture distribution $G_{\mathbf{x}}$, allows for accurate detection of rating scale category usage in the posterior distribution of $G_{\mathbf{x}}(\cdot)$ for covariates $\mathbf{x}$ of interest, for example, in the posterior means $\mathrm{E}_n[G_{\mathbf{x}}(\cdot)]$. This could help reveal when subsets or all category labels are unclear, or when DIF is present.

Another unique feature of the DDP-RM is that it clusters item category thresholds based on the similarity in the mixing distribution. This similarity is captured through the neighborhood inducing parameter $\boldsymbol{\gamma}$. When two separate γs have the same values, the mixture components are the same for the covariates associates with the two γs. Then, similar γs would indicate that the items associated with the γs have similar mixing distributions for the rating category thresholds.

Illustration of the DDP-RM on Simulated Data

In this section, the authors provide a simulation study to demonstrate the model's ability to correctly identify DIF due to the presence of multiple latent examinee clusters and the item free of DIF.

The authors generated item response data for 3,000 examinees and 10 items, with each item scored on a 0 to 2 rating scale, yielding a total of $n = 30{,}000 = 3{,}000 \times 10$ rating observations. These data were generated according to the parameters of a two-mixture Rasch logistic rating scale model, which are described as follows. Each simulated examinee was assigned an ability θ parameter according to an independent draw from a normal $\mathrm{n}(0, 2.25)$ distribution. In addition, each examinee was randomly assigned to one of two clusters, with equal probability. As a result, 1,505 and 1,496 examinees were assigned to the first and second cluster, respectively. Furthermore, each of the first 9 items was specified to be free of DIF in the rating category thresholds, with the second threshold parameter (τ_2) being 1 unit larger than the first threshold parameter (τ_1). For example, the fifth item was assigned thresholds $\boldsymbol{\tau} = (\tau_1 = -.5, \tau_2 = .5)^{\mathrm{T}}$. Over these 9 items, the category thresholds had range $(-2.3, 2.3)$. In contrast, the 10th item was specified to have DIF at the threshold parameter τ_2, but not at threshold τ_1. Specifically, for this item, the first threshold was specified as $\tau_1 = -1.25$ for both examinee clusters. The

second threshold parameter was specified as $\tau_2 = 0$ for the first examinee cluster and $\tau_2 = 2$ for the second examinee cluster.

To analyze the simulated rating data using the DDP-RM, the authors made the following model specifications for the purposes of demonstrating the model's ability to differentiate between DIF and DIF-free items. First, they treated only 2 items as having random (mixed) threshold parameters. They included the 5th item, which was free of DIF, and the 10th item, which had DIF. For each of the remaining 8 items, the thresholds were treated as fixed (nonmixed) parameters. In addition, for the model, the authors specified covariates $\mathbf{x}$ as 0-1 dummy indicators of the 10 items. Thus, the authors wrote the neighborhood size parameter as $\psi(\mathbf{x}) = \psi_j$. Furthermore, they assigned proper prior distributions to the model's parameters, namely, $\theta_t \sim_{iid} \text{n}(0, \sigma^2)$, $\sigma^2 \sim \text{ig}(1,1)$, $\boldsymbol{\tau}_h \sim_{iid} \text{n}(\mathbf{0}, 2\mathbf{I}_m)$, $\upsilon_h \sim_{iid} \text{beta}(1, \alpha)$, $\alpha \sim \text{ga}(1,1)$, $\gamma_j \sim_{iid} \text{un}(1, 745)$, while fixing $\psi_j = 5$ for all items they treated as random. For each of the 8 items with fixed (nonmixed) threshold parameters, the thresholds were assigned prior $\boldsymbol{\tau} \sim \text{n}(0, 10\mathbf{I}_m)$. Such prior distributions may be specified for typical real-data applications of the DDP-RM, where little prior information is available about the model parameters.

In order to perform Bayesian posterior estimation of the DDP-RM parameters, the authors ran the MCMC sampling algorithm for 200,000 MCMC sampling iterations. They discarded the first 100,000 MCMC samples (i.e., burn-in period), and saved every fifth sample thereafter, for a total of 20,000 MCMC samples that they saved and used for posterior inference. According to standard convergence diagnostics (Geyer, 2011), these samples showed good mixing of model parameters, and the posterior estimates of these parameters typically had 95% MCMC confidence intervals of around .01.

For the DDP-RM, the posterior mean estimates of the mixing distribution $G_{\mathbf{x}}(\boldsymbol{\tau})$, given covariates $\mathbf{x}$ (e.g., item indicators), reveal how examinees used the rating categories. For the fifth item, the top two panels of Figure 1 present the (marginal) posterior mean density estimates of the mixture distributions $G_{\mathbf{x}}(\tau_1)$ and $G_{\mathbf{x}}(\tau_2)$, which correspond to the two rating threshold parameters. As shown in the figure, for each of the two thresholds of this fifth item, the marginal posterior mean density estimate was unimodal with a very small variance. Thus, these estimates correctly show that the item is free from DIF in that a single common set of category thresholds applies to all examinees. That is, there is a single cluster of examinees in terms of these thresholds. Moreover, the posterior mean estimates of the thresholds were $\bar{\boldsymbol{\tau}} = (\bar{\tau}_1 = -.44, \bar{\tau}_2 = .43)^{\text{T}}$, and are thus very similar to the true data-generating values of $\boldsymbol{\tau} = (\tau_1 = -.5, \tau_2 = .5)^{\text{T}}$.

The bottom two panels of Figure 1 contain the estimated (marginal) posterior densities of $G_{\mathbf{x}}(\tau_1)$ and of $G_{\mathbf{x}}(\tau_2)$ for the two rating threshold parameters associated with the 10th item. For this item, the estimated marginal posterior density of the first threshold $G_{\mathbf{x}}(\tau_1)$ is unimodal. Thus, this estimate correctly indicates that the threshold parameter τ_1 is DIF free. The marginal posterior density estimate of the second threshold, however, is bimodal. Hence, this estimate correctly indicates that there is DIF for that item in that threshold. In other words, there are two latent clusters (modes) of examinees in terms of that threshold parameter. Furthermore, the first mode is slightly less than 0, and the second mode is approximately 2. These values are close to the generating values (0 and 2, respectively) that were used to simulate the rating data.

Illustration of the DDP-RM on Real Data

In this section, the authors illustrate the DDP-RM through the analysis of a real data set obtained from the verbal aggression study, which was based on the Verbal Aggression questionnaire (De Boeck & Wilson, 2004). Moreover, they compare the predictive performance between the DDP-RM and several other IRT rating models. Specifically, this data set contains ratings of 24 items that were made by each of 316 students (243 females and 73 males) who

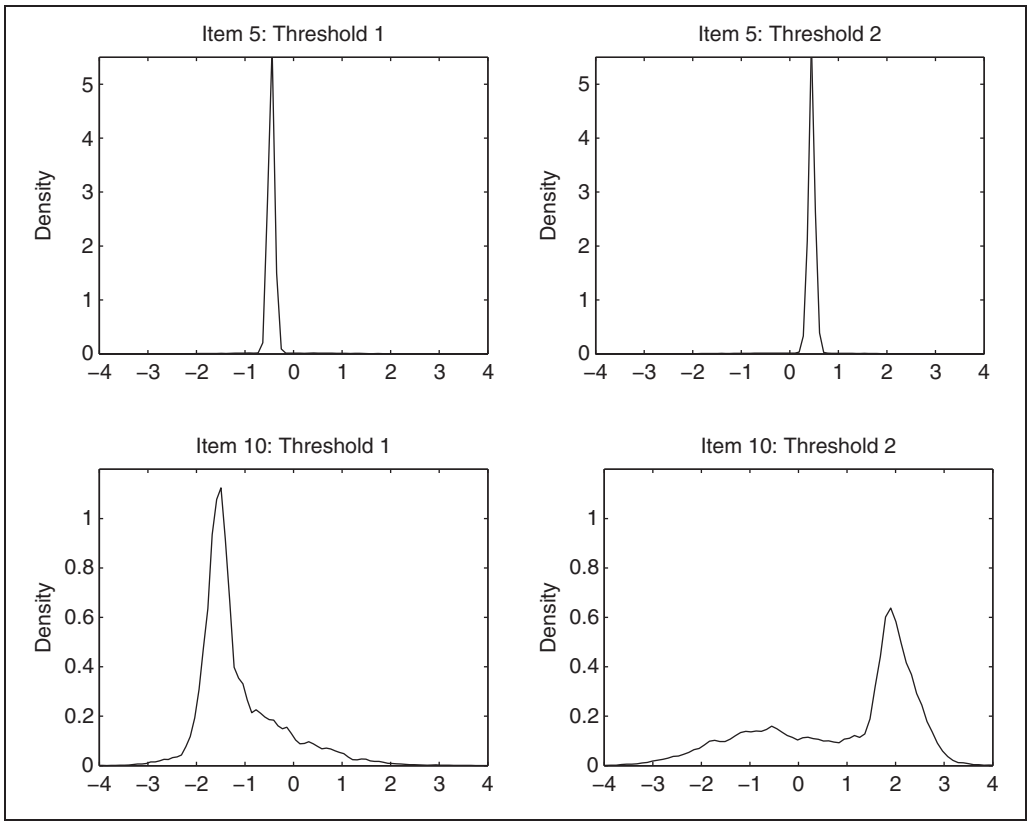

Figure 1. For the simulated data, the marginal posterior mean density estimates of the rating category thresholds, for two items.

attended a Dutch-speaking Belgian university. Each of the 24 items of the Verbal Aggression questionnaire represents a type of verbal aggression (e.g., ''A bus fails to stop for me. I would want to curse.''), and can be categorized into a $2 \times 2 \times 3$ design: Behavior Mode (Want or Do) by Situation Type (Other-to-blame or Self-to-blame) by Behavior Type (Curse, Scold, or Shout). Each item was scored on a rating scale of 0 = *no*, 1 = *perhaps*, and 2 = *yes*.

To analyze the verbal aggression rating data using the DDP-RM, the authors treated all items as having random (mixed) threshold parameters. Also, as before, they specified the covariates $\mathbf{x}$ as 0-1 dummy indicators for the 24 Verbal Aggression items. Hence, the authors could rewrite the neighborhood size parameter as $\psi(\mathbf{x}) = \psi_j$. Furthermore, they assigned priors $\theta_t \sim_{iid} \mathrm{n}(0, 1)$, $\boldsymbol{\tau}_h \sim_{iid} \mathrm{n}(0, 5\mathrm{I}_m)$, $\upsilon_h \sim_{iid} \mathrm{beta}(1, \alpha)$, $\alpha \sim \mathrm{ga}(1, 1)$, $\boldsymbol{\gamma}_j \sim_{iid} \mathrm{un}(1, 745)$, and $\psi_j \sim_{iid} \mathrm{un}(.5, 20)$, in their attempt to specify rather noninformative priors for the model parameters. Finally, as is done with other IRT models, they assumed that the item responses of the Verbal Aggression questionnaire are independent, conditionally on all model parameters. As each of the 24 questionnaire items can be classified according to $2 \times 2 \times 3$ design in terms of item type, there may be a concern that the data violate this assumption. If such a concern arises, then one can specify additional covariates in the DDP-RM that describe the levels of this design, so that it becomes more reasonable to assume conditional independence under the (expanded) model. However, for the interests of providing a simple illustration of the DDP-RM, the authors will analyze the data by specifying the covariates $\mathbf{x}$ as 0-1 dummy indicators of the 24 questionnaire items.

To perform Bayesian posterior estimation of the DDP-RM parameters, the authors ran the MCMC sampling algorithm for 200,000 MCMC sampling iterations. They discarded the first 100,000 MCMC samples (i.e., burn-in period), and saved every fifth sample thereafter, for a total of 20,000 MCMC samples that they saved and used for posterior inference. According to standard convergence diagnostics (Geyer, 2011), these samples displayed good mixing of model parameters, and the posterior estimates of these parameters typically had 95% MCMC confidence intervals of around .01.

For three of the Verbal Aggression questionnaire items, Figure 2 contains the marginal posterior mean density estimates of $G_{\mathbf{x}}(\tau_1)$ and $G_{\mathbf{x}}(\tau_2)$ for the two rating threshold parameters. As shown, Items 1 and 23 exhibit greater variability in their rating category thresholds compared with Item 2. For Item 1, the marginal posterior mean density estimate of the first threshold (τ_1) and the second threshold (τ_2) is tri-modal and bimodal, respectively. Thus, the item contains DIF in the sense that there are three distinct latent clusters of examinees with respect to threshold τ_1, and two distinct latent clusters of examinees with respect to threshold τ_2. For Item 23, the marginal posterior mean density estimate is bimodal for threshold parameter τ_1 and for threshold parameter τ_2. Hence, this item also contains DIF. However, for Item 2, the marginal posterior mean density estimate of each threshold is unimodal with small variance. Thus, these estimates suggest the lack of DIF and indicate that there is a single cluster of examines in terms of these threshold parameters.

In total, 21 of the 24 items were found to be unimodal. The multimodal items, such as Item 23, can be further examined for the causes of DIF, and may lead to insights into ways of modifying this questionnaire item. However, the authors need not remove such items because the DDP-RM produces posterior parameter estimates (e.g., of examinee ability parameters) after controlling for any DIF. In contrast, an IRT model that assumes no DIF may provide misleading parameter estimates, when this assumption is empirically violated.

For all 24 Verbal Aggression questionnaire items, the authors present in the Table of Online Appendix B the marginal posterior means, standard deviations, and modes of the threshold distributions $G_{\mathbf{x}}(\tau_1)$ and $G_{\mathbf{x}}(\tau_2)$. The posterior means for the thresholds ranged from -0.68 to 3.32 across the items. Similar to conclusions by others (e.g., De Boeck & Wilson, 2004), Item 21 was found to be the most difficult to endorse, as it attained the largest posterior means for the category thresholds. Item 4 was the easiest to endorse, as it had the smallest posterior means for the category thresholds. In addition, over all the 24 items, the marginal posterior mean estimates of the neighborhood location parameters γ_j ranged from 6.0 to 255.6, whereas the marginal posterior mean estimates of neighborhood size parameters ψ_j ranged from 7.5 to 19.8. In terms of the posterior means, the items had noticeably different neighborhood locations and sizes, indicating that the items differed in terms of the mixing distribution $G_{\mathbf{x}}(\boldsymbol{\tau})$. Finally, over the 316 examinees (students), the marginal posterior mean estimates of the ability parameters θ_t had range $(-2.37, 3.74)$, along with a mean -0.02 and standard deviation 1.01.

The authors then compared the predictive performance between the DDP-RM, and other well-known IRT rating models. They did so on the basis of the $D(\underline{m})$ model selection criterion (Gelfand & Ghosh, 1998). For each model, the criterion is defined by its posterior predictive mean-square error. A detailed description of the criterion is provided in Online Appendix C. The authors found that the DDP-RM (with $D(\underline{m}) = 4{,}984$) outperformed seven other IRT models, by at least 49 $D(\underline{m})$ units. The other models include the PCM ($D(\underline{m}) = 5{,}716$), the generalized PCM ($D(\underline{m}) = 5{,}686$), the RSM ($D(\underline{m}) = 5{,}726$), the GRM ($D(\underline{m}) = 5{,}709$), the nominal response model ($D(\underline{m}) = 5{,}689$; Bock, 1972), a three-mixture PCM ($D(\underline{m}) = 5{,}163$) (with the optimal number of mixture components determined by the Akaike Information Criterion; Akaike, 1973; Rost, 1991), and a covariate-independent DP mixture PCM model that treated the category thresholds as random ($D(\underline{m}) = 5{,}033$). The software IRTPRO 2.1 (Cai, Thissen, & du Toit, 2011)

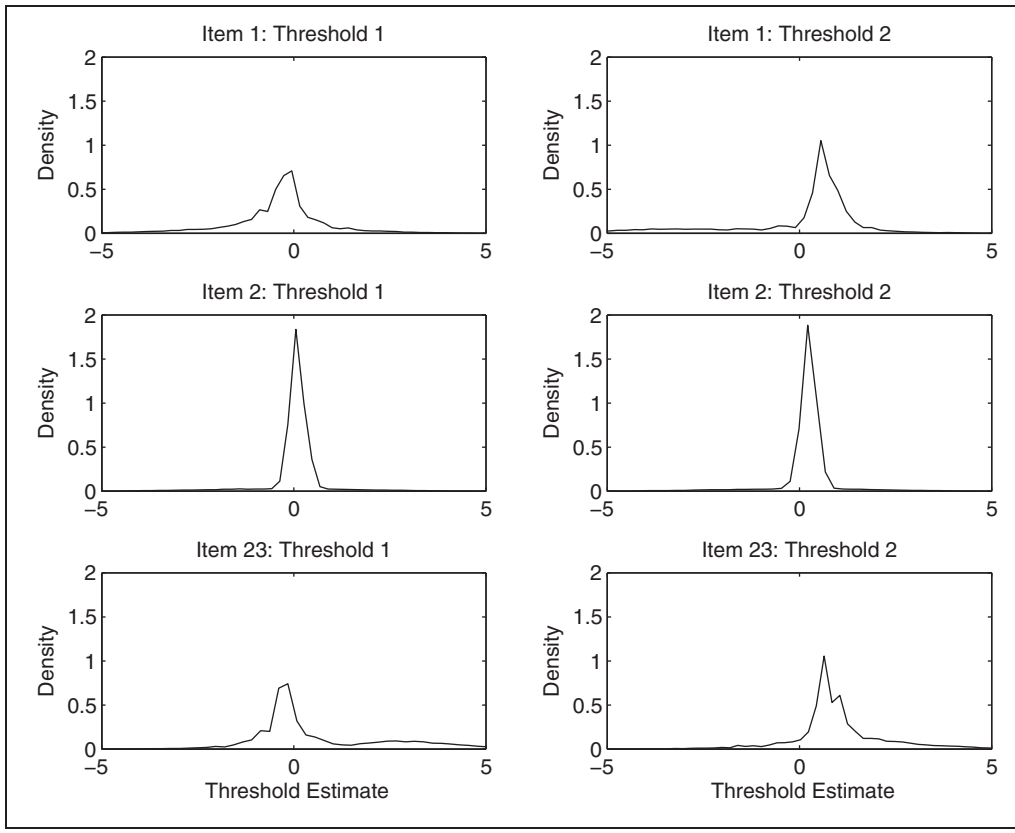

Figure 2. For three items of the Verbal Aggression questionnaire, the marginal posterior mean density estimate of the rating category thresholds.

and WINMIRA 2001 (von Davier, 2001) were used to fit the other models. The DP mixture PCM model was estimated on the basis of 20,000 converged MCMC posterior samples, using MATLAB code (2012, The MathWorks, Natick, Massachusetts). In all, the three-mixture models outperformed the traditional, nonmixture models. Nevertheless, these results suggest that allowing the mixing distribution to vary across items leads to positive gains in fit.

Conclusion

The authors introduced the DDP-RM, a novel Bayesian nonparametric rating scale IRT model, which is an infinite-mixture model based on the local DP. They showed that the model, through posterior mean estimates of the mixing distribution for the threshold parameters, describes how the examinees used the rating categories. The posterior number of modes in the mixing distribution reveals the number of clusters (groups) of examinees in terms of item category thresholds. Moreover, using real rating data, the authors demonstrated that the new model provides a substantially better predictive fit of the rating data compared with other IRT models commonly used.

In future research, the DDP-RM can be extended by assigning a nonparametric prior for the ability distribution, such as a DP prior (San Martín et al., 2011). In addition, it would be of interest to extend the model by specifying $G_{\mathbf{x}}(\boldsymbol{\tau})$ with a more flexible, infinite mixture, such as

mixture weights based on an infinite-ordered probits regression model (Karabatsos & Walker, 2012).

Acknowledgments

The authors thank Professor Stephen G. Walker, University of Texas at Austin, for helpful conversations about the Markov chain Monte Carlo (MCMC) algorithm they used for the article. In addition, the authors thank two anonymous reviewers for their comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by NSF Research Grant SES-1156372, from the Program in Methodology, Measurement, and Statistics.

Supplemental Material

The online appendices are available at http://apm.sagepub.com/supplemental

References

Akaike, H. (1973). Information theory and the extension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest, Hungary: Academiai Kiado.

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Chicago, IL: Scientific Software International.

Chung, Y., & Dunson, D. B. (2011). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, *63*, 59-80.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*, 41-66.

Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software*, *48*, 1-25.

Gelfand, A. E., & Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, *85*, 1-11.

Geyer, C. (2011). Introduction to MCMC. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 3-48). Boca Raton, FL: CRC Press.

Kalli, M., Griffin, J. E., & Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, *21*, 93-105.

Karabatsos, G., & Walker, S. G. (2012). Adaptive-modal Bayesian nonparametric regression. *Electronic Journal of Statistics*, *6*, 2038-2068.

Karabatsos, G., & Walker, S. G. (in press). Bayesian nonparametric IRT. In W. van der Linden & R. Hambleton (Eds.), *Handbook of item response theory: Models, statistical tools, and applications*. New York, NY: Taylor & Francis.

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

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science* (pp. 50-55). Alexandria, VA: American Statistical Association.

MacEachern, S. N. (2000). *Dependent Dirichlet processes*. Unpublished manuscript, Department of Statistics, The Ohio State University, Columbus.

MacEachern, S. N. (2001). Decision theoretic aspects of dependent nonparametric processes. In E. George (Ed.), *Bayesian methods with applications to science, policy and official statistics* (pp. 551-560). Creta, Greece: International Society for Bayesian Analysis.

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

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley-Interscience.

Miyazaki, K., & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika*, *74*, 375-393.

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

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75-92.

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

Samejima, F. (1972). *A general model for free response data* (Psychometrika Monograph No. 18). Richmond, VA: Psychometric Society.

San Martín, E., Jara, A., Rolin, J. M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, *76*, 385-409.

Seol, H. (1999). Detecting differential item functioning with five standardized item-fit indices in the Rasch model. *Journal of Outcome Measurement*, *3*, 233-247.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639-650.

Smit, A., Kelderman, H., & van der Flier, H. (2003). Latent trait latent class analysis of an Eysenck Personality Questionnaire. *Methods of Psychological Research Online*, *8*, 23-50.

Smith, R. M., & Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement*, *4*, 153-163.

von Davier, M. (2001). WINMIRA 2001 [Computer software]. St. Paul, MN: Assessment Systems Corporation.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*, 389-406.

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.