Dependent Dirichlet Process Mixture Modeling

of Rating Category Usage

BY

KEN AKIRA FUJIMOTO
B.A., California State University, Long Beach, 1997
M.F.A., University of Arizona, 2004

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, IL

Defense Committee:

George Karabatsos, Chair and Advisor
Everett V. Smith, Jr., Educational Psychology
Donald Hedeker, Biostatistics
Ryan Martin, Mathematics, Statistics, and Computer Sciences
Stanley L. Sclove, Information and Decision Sciences

This thesis is dedicated to my mother, father, Claire, and Emma. At an early age, my parents instilled in me the value of education and the importance of never giving up. This thesis is directly tied to these lessons. My wife, Claire, has supported me and my academic endeavors since the day we met. My daughter, Emma, has sacrificed playtime with me in order for me to finish this thesis. Without the support and understanding of my wife and daughter, this thesis would not have been possible.

ACKNOWLEDGEMENTS

I am indebted to my advisor, George Karabatsos, for his time and mentorship over the years. Without his patience and the countless number of hours he spent discussing nonparametric statistics with me, this dissertation would not have been possible. He has helped me reach a level in my research that I did not think was possible when I first began my graduate career. I also express gratitude to the other members of my dissertation committee, Ev Smith Jr., Don Hedeker, Ryan Martin, and Stan Sclove. Their comments and suggestions have helped advance this dissertation. I also express thanks to Rachel Gordon and Wei-Chin Hwang. My work with them gave me insight into some of the challenges in rating scale data analysis, one of which inspired this dissertation.

Last but not least, I want to show my appreciation to my wife, Claire. She endured with me the ups and downs during the dissertation process. Throughout, her support never wavered, even when the process took longer than expected.

Finally, I note that this dissertation research was supported by NSF-MMS Research Grant SES-1156372.

KAF

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

FIGURE PAGE

LIST OF ABBREVIATIONS

AFD	Acculturative Family Distancing
AIC	Akaike Information Criteria
AP	Advanced Placement
BIC	Bayesian Information Criteria
CDF	Cumulative Distribution Function
CMLE	Conditional Maximum Likelihood Estimation
CORE	Common Odds Ratio Estimator
DIC	Deviance Information Criterion
DDP	Dependent Dirichlet Process
DDP-RM	Dependent Dirichlet Process Rating Model
DIF	Differential Item Functioning
DP	Dirichlet Process
DSF	Differential Step Functioning
EMD	Equal-Mean-Difficulty
GF	Goodness-Of-Fit
GLMM	Generalized Linear Mixed Models
GPCM	Generalized Partial Credit Model
GRM	Graded Response Model
IRT	Item Response Theory
ISRF	Item Step Response Function
lDP	Local Dirichlet Process

LIST OF ABBREVIATIONS (continued)

LOOCV	Leave-One-Out Cross Validation
LPML	Log Predicted Marginal Likelihood
LRT	Likelihood Ratio Test
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
mDP	Multiple Dirichlet Process
MFM	Model Fit Measures
MGRM	Modified Graded Response Model
MHM	Monotone Homogeneity Model
MMLE	Marginal Maximum Likelihood Estimate
NRM	Nominal Response Model
PCM	Partial Credit Model
Pen	Penalty
PMF	Probability Mass Function
RMSD	Root Mean Squared Deviation
RSM	Rating Scale Model
SD	Standard Deviation

SUMMARY

Educational and psychological tests are often utilized to measure latent constructs, such as math achievement or self-esteem, in a sample of persons. Comparisons of subgroups within the sample are made with respect to test scores. An underlying assumption when such comparisons are made is that the item scores have the same meaning across the subgroups under comparison. That is, the person characteristics that distinguish the subgroups (e.g., gender and race) are not part of the response process. Unfortunately, this assumption does not always hold and should be tested. What makes testing this assumption more challenging when the items are of rating type (i.e., polytomously scored) is that the person characteristics could have differential effects on the response process across the scores. That is, for some subset of scores for an item, the response process could be free of such effect while for another subset of scores for the same item, such effect could be part of the response process.

A differential step functioning (DSF) analysis can determine whether person characteristics are part of the response process in rating scale items, and if so, whether person characteristics have an equal or differential effect across all response categories. Item response theory (IRT) models are convenient tools for performing a DSF analysis. The traditional approach using multiple-group IRT includes the person characteristics of interest into the model. This approach, however, has its limitations. It requires some method of linking across the subgroups, with the choice having potential consequences on the effectiveness of the model detecting DSF. Additionally, it cannot account for when the subgroups are latent (i.e., latent classes). A finite-mixture IRT model can examine for whether DSF occurs across latent classes. Unfortunately, it also has its limitations. It assumes that the same number of mixture components describe the data for all items.

For this dissertation, I introduce a Bayesian nonparametric IRT model, based on covariate-dependent infinite-mixture modeling, to address these limitations of multiple-group and finite-mixture IRT models. The mixing distribution for this model is formed using the multiple Dirichlet Process (mDP), which is a type of dependent Dirichlet Process, and this distribution is allowed to flexibly vary across items. Two simulation studies and analyses of two real-life rating data sets indicated that DSF across latent classes is revealed in the shape of the posterior mean estimates of the mixing distributions produced with the mDP model. When an item is free of DSF, the posterior density corresponding to each category step is unimodal with small variance. When an item has DSF, the posterior density corresponding to the category step where the DSF resides is multimodal, with the number of modes indicating the number of latent classes contributing to the DSF. The simulation studies also indicated that sample size and the magnitude of the DSF influence the effectiveness of the mDP model's ability to detect DSF.

The results of the simulation studies show that, when an appropriate sample size and DSF magnitude are present, the mDP model provides a unique approach to identifying where the DSF resides in rating scale items. The DSF is displayed visually through the posterior densities of the mixing distributions, and the mDP model accomplishes this while addressing the limitations of the traditional IRT approaches to DSF analysis.

I. LITERATURE REVIEW

Introduction to Item Response Modeling of Rating Category Usage

A test is broadly considered to be an assessment instrument that contains a set of items. It is administered to a sample of persons to gather information about their achievement or attitude score on some latent construct. When a test functions differently across two or more groups of persons, it is difficult to compare scores across the groups based on a common frame of reference (Berk, 1982; Camilli & Shepard, 1994). This differential functioning is caused by the presence of construct-irrelevant variance, defined as the variance in the test scores due to factors (e.g., group membership) unassociated with the construct (i.e., the latent trait) that the scores are intended to represent. The presence of construct-irrelevant variance should be addressed, because it suggests that factors other than the construct of interest affect persons' responses to items, which in turn, threaten construct validity (Kane, 2006; Messick, 1989, 1995). Specifically, construct-irrelevant variance threatens the internal structural aspect of construct validity; that is, the extent to which items contained in the rating instrument and the construct for which the items are indicators are related (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

Such differential test functioning can negatively affect many different fields. For example, if an advanced placement (AP) test functions differently across groups of students, then one unqualified student could receive a score that qualifies her or him for placement into an AP course, whereas a qualified student could receive a score that denies her or him access to such higher-level course. In the setting of the health and social sciences, conclusions based on data gathered from noninvariant tests can be inaccurate and misleading, possibly resulting in

consequences such as misguided policy initiatives. In a clinical setting, incorrect diagnoses can result if noninvariant instruments are used to diagnose persons according to a single cutoff score.

An important source of construct-irrelevant variance is differential item functioning (DIF). An item is said to have DIF when the probability of an item response is not constant across all person subgroups after controlling for a latent trait (Angoff, 1993; Holland & Thayer, 1988). For rating items, a necessary condition for lack of DIF is invariance of the rating category parameters across all subgroups of the person population. An item that lacks such invariance also contains differential step functioning (DSF). If all items are free of DIF and DSF, then support exists for the internal aspect of construct validity.

Typically, there are two approaches to dealing with the presence of DIF or DSF. One is to first statistically test items for the presence of DIF or DSF, and then either remove any problematic items before calculating scores, or take into consideration the extent of DIF or DSF when interpreting the scores (e.g., AERA et al., 1999; Camilli & Shepard, 1994; Holland & Thayer, 1988; Penfield & Camilli, 2007). The second approach is to control for the presence of DIF or DSF in a statistical model when estimating a person's score (i.e., the person's latent trait that the test is designed to measure) (e.g., Chu & Kamata, 2007; De Jong & Steenkamp, 2010; De Jong, Steenkamp, & Fox, 2007; Sireci, 2005).

The remainder of this chapter provides a review of the literature on item response theory (IRT) for rating scale data and its connection with generalized linear mixed models (GLMM), the formal definitions of DIF and DSF, the statistical models employed for DSF analysis on rating scale data that do or do not produce adjusted scores, and an existing model that can be used to investigate DSF. The review concludes with a short discussion of potential research directions with respect to the current problems in DSF research.

Review of Polytomous IRT and Generalized Linear Mixed Models

An IRT model parameterizes both latent trait levels from the persons' responses and the items to which the persons provide responses (Embretson & Reise, 2000). There are many different IRT models, each characterized by a set of assumptions. In general, any given IRT model assumes that the joint probability of a person's responses to items on a test is expressed as:

$$P\left(Y_1 = y_1,...,Y_j = y_j,\ldots,Y_J = y_J\right) = \int \prod_{j=1}^{J} P\left(Y_j = y_j \mid \theta;\boldsymbol{\beta}\right) \mathrm{d}G\left(\theta\right),$$

where P denotes probability; θ is a real-valued (possibly multidimensional) parameter representing the latent trait for a respondent; $\boldsymbol{\beta}$ is the vector of item and response-category parameters; and $G(\theta) = P(\Theta \leq \theta)$ is the population distribution of latent trait levels, where $\Theta = (-\infty, \theta]$ is a bin of subsets of values from the sample space that are less than or equal to θ (Holland & Rosenbaum, 1986).

In general, IRT models make at least three assumptions: (a) the person latent trait, θ, is real-valued, (b) all item responses have local independence, meaning that they are conditionally independent given θ; that is:

$$P\left(Y_1 = y_1,...,Y_j = y_j,\ldots,Y_J = y_J \mid \theta;\beta\right) = \prod_{j=1}^{J} P\left(Y_j = y_j \mid \theta;\boldsymbol{\beta}\right),$$

and (c) the item step response function (ISRF) is monotonically increasing, which means that the probability of Y_j being greater than or equal to k, which is formally represented by $P\left(Y_j \geq k \mid \theta;\boldsymbol{\beta}\right)$, is nondecreasing in each coordinate of θ for all possible ordered categories for the item, given by $k = 0, 1, \ldots, m_j$, where m_j represents the maximum score category for item j. This states that the probability of Y_j being greater than or equal to k never decreases as the latent trait level increases. Rasch models (e.g., Rasch, 1960) and the double monotonicity model

(Mokken & Lewis, 1982) make a further assumption of invariant item ordering—that it is possible to assign indices $i = 1, 2, \ldots, J$ to all items of the test, such that:

$$E\left(Y_1 \mid \theta\right) \leq E\left(Y_2 \mid \theta\right) \leq \ldots \leq E\left(Y_i \mid \theta\right) \leq E\left(Y_J \mid \theta\right), \text{ for all values of } \theta.$$

Usually, an assumption about the form of the distribution of the population ability $G(\theta)$ is specified beforehand; for example, as a normal distribution (i.e., $G[\theta] = \text{Normal}[\mu_\theta, \tau_\theta]$) with a mean of zero (i.e., $\mu_\theta = 0$) and variance either fixed at some value, such as 1 (i.e., $\tau_\theta = 1$), or an unknown variance (τ_θ) to be estimated from the data at hand. The distribution $G(\theta)$ can be more flexibly modeled with a finite number of mixtures. That is, $G(\theta) = \sum_{h=1}^{H} \text{normal}\left(\theta \mid \mu_h, \sigma_h^2\right) w_h$, with the H number of mixture weights following a multinomial distribution $w_1, \ldots, w_H \sim \text{multinomial}\left(\boldsymbol{\psi}, 1\right)$, with $\boldsymbol{\psi} = \left(\psi_1, \ldots, \psi_H\right)$ and $\sum_{l=1}^{L} \psi_l = 1$. An even more flexible way to model the distribution $G(\theta)$ is to model it nonparametrically so that no prior assumption is made on the specific form of the distribution. Moreover, the population distribution of latent trait levels $G(\theta)$ is assumed to be either continuous or discrete, and an IRT model under a discrete distribution assumption for $G(\theta)$ is referred to as a latent class IRT model. An important example of a nonparametric model for G, which gives rise to a discrete infinite-mixture IRT model, is provided by the Dirichlet Process (DP) model, parameterized by (α, G_0), the precision parameter and the mean (baseline) parameter of $G(\theta)$, respectively. Specifically, a random $G(\theta)$ from the DP is constructed by taking $G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}\left(\cdot\right)$, with mixture weights $w_h = \upsilon_h \prod_{l<h} \left(1 - \upsilon_l\right)$, $\upsilon_1, \upsilon_2, \ldots \underset{iid}{\sim} \text{Beta}\left(1, \alpha\right)$, and $\theta_1, \theta_2, \ldots \underset{iid}{\sim} G_0$, for $h = 1,2,\ldots$, where $\delta_{\theta_h}\left(\cdot\right)$ denotes the degenerate distribution that assigns probability mass 1 on the value θ_h (Sethuraman, 1994).

Unidimensional rating data IRT models, in which the assumption of local item independence and monotonicity are also met, achieve a weaker form of stochastic ordering of the latent trait (SOL). That is, let θ^* represent an arbitrary point on the latent trait scale, then

$$P\left(\theta \geq \theta^* \mid X_+ \geq X^*\right) \geq P\left(\theta \geq \theta^* \mid X_+ < X^*\right)$$

for all θ^* and all total test scores X^*. This states that stochastic ordering of the expected latent trait is achieved for high and low total test score, X_+, and groups of persons (with a total test score X^* distinguishing the two groups), which justifies ordering persons on the latent trait based on the total score (van der Ark & Bergsma, 2010).

For rating scale data, there are two general classes of IRT models, which have had many applications. One class consists of cumulative probability models, which model the probability of endorsing category k (where $k = 0, 1, \ldots, m_j$, with $m_j > 1$) on item j given latent trait level and item parameters by:

$$P\left(Y_j = k \mid \theta; \boldsymbol{\beta}\right) = \begin{cases} H_{jk}\left(\eta\right), \text{when } k = m_j \\ H_{jk}\left(\eta\right) - H_{j(k+1)}\left(\eta\right), \text{when } k \neq 0 \text{ or } m_j, \\ 1 - \sum_{k=1}^{m_j} P\left(Y_j = k \mid \theta; \boldsymbol{\beta}\right), \text{when } k = 0 \end{cases} \tag{1.1}$$

where $H\left(\cdot\right)$ is the cumulative distribution function (CDF), and $H\left(\eta\right) = P\left(\vartheta \leq \eta\right)$, with $\vartheta = \left(-\infty, \eta\right]$ a bin of a subset of values from the sample space that are less than or equal to η. The CDF, $H_{jk}\left(\eta\right)$, represents an ISRF, which governs the probability of a Y_j being greater than or equal to k based on the systematic component η, which is a linear combination of the parameters θ and β. That is, $H_{jk}\left(\eta\right) = P\left(Y_j \geq k \mid \theta, \beta\right)$.

The other class of IRT models is defined by adjacent category probabilities, which model the probability of endorsing category *k* on item *j* given latent trait level and item parameters by

$$P\left(Y_j = k \mid \theta; \boldsymbol{\beta}\right) = \frac{\prod_{x=0}^{k} H_{jx}\left(\eta\right) \prod_{l=k+1}^{m_j} \left[1 - H_{jl}\left(\eta\right)\right]}{\sum_{w=0}^{m_j} \left(\prod_{x=0}^{w} H_{jx}\left(\eta\right) \prod_{l=w+1}^{m_j} \left[1 - H_{jl}\left(\eta\right)\right] \right)}, \text{ with } \prod_{x=0}^{0} H_{jx}\left(\eta\right) \equiv 1. \qquad (1.2)$$

In this case, the CDF $H_{jk}\left(\eta\right)$ represents the probability of a response of *k* given the person ability, the item category steps, and that the category options are limited to *k* and *k* – 1 for item *j*. That is, $H_{jk}\left(\eta\right) = P\left(Y_j = k \mid \theta, \beta, k-1, k\right)$

Using the terminology of generalized linear models (e.g., McCullagh & Nelder, 1989), the cumulative distribution function $H\left(\eta\right) : \mathbb{R} \rightarrow \left(0,1\right)$ is the inverse-link function with systematic component η having space $\Re$, corresponding to link $h\left(\cdot\right) = H^{-1}\left(\cdot\right)$, with range $\Re$. An assumption is usually made about the form of the cumulative distribution for $H\left(\cdot\right)$, and two common types are the logistic and normal ogive. When the CDF is the logistic distribution with a mean of zero and scale of 1, then

$$H\left(\cdot\right) \equiv L\left(\cdot \mid 0,1\right) = \frac{\exp\left(\cdot\right)}{1 + \exp\left(\cdot\right)},$$

and when the CDF is the normal ogive with a mean of 0 and scale of 1, then

$$H\left(\cdot\right) \equiv N\left(\cdot \mid 0,1\right) = \Phi\left(\cdot\right).$$

The inverse link, $H\left(\cdot\right)$, can also be modeled nonparametrically; the only constraint is that the distribution is nondecreasing in each coordinate of θ for all *k* (e.g., Miyazaki & Hoshino, 2009; Newton, Czado, & Chappell, 1996; Sijtsma & Molenaar, 2002). For example, it is possible to

model $H(\cdot)$ with a DP model that supports the entire space of (measurable) cumulative distribution functions on $\Re$ (e.g., Newton et al., 1996), or with an isotonic regression model that assumes that the expected rating response is nondecreasing with some estimate of a unidimensional latent trait θ, such as a total test score (e.g., Karabatsos & Sheu, 2004).

Examples of Common IRT Models for Rating Data

In this section, I present the systematic components for cumulative and adjacent category probability IRT models commonly used to analyze rating data. For information about other types of common IRT models, refer to Table I.

Cumulative probability models. One type of cumulative probability IRT model is the graded response model (Samejima, 1969), which has the systematic component

$$\eta_{jk} = \alpha_j(\theta - \delta_{jk}), \text{ with the constraint } \delta_{j1} < \delta_{j2} < \ldots < \delta_{jm_j},$$

where $\alpha_j > 0$ represents the discrimination parameter for item j, and δ_{jk} is item j's step parameter for category k, sometimes decomposed as $\delta_{jk} = \beta_j + \tau_{jk}$, where β_j is the overall item difficulty level for item j, and τ_{jk} is the item's relative step parameter for category k. A special case of the GRM is the modified graded response model (Muraki, 1990), which assumes that all items share a common set of category steps; that is, $\delta_{jk} = \beta_j + \tau_k$. The M-GRM and GRM, with the relative-step parameterization, require an additional constraint to be placed on the τ_ks for model identification. One such constraint is that the sum of the step estimates must be zero. Another way to place restrictions on the τ_ks is to assign a prior distribution that supports the order constraint $-\infty \equiv \tau_0 < 0 \equiv \tau_1 < \tau_2 < \ldots < \tau_{m_j+1} \equiv \infty$.

TABLE I

SELECTED SYSTEMATIC COMPONENTS FOR *H* TO REPRESENT DIFFERENT IRT MODELS

Model	Systematic Component (η)	$P\left(Y_j = k \mid \theta; \boldsymbol{\beta}\right)$
Modified Graded Response Model (M-GRM) (Muraki, 1990)	$\alpha_j\left(\theta - \beta_j - \tau_k\right)$, with $\tau_1 < \tau_2 < \ldots < \tau_m$	$H\left(\eta_{jk}\right) - H\left(\eta_{j(k+1)}\right)$
Graded Response Model (GRM) (Samejima, 1969)	$\alpha_j\left(\theta - \delta_{jk}\right)$, with $\delta_{j1} < \delta_{j2} < \ldots < \delta_{jm_j}$	$H\left(\eta_{jk}\right) - H\left(\eta_{j(k+1)}\right)$
Rasch Rating Scale Model (RSM) (Andrich, 1978)	$\theta - \left(\beta_j + \tau_k\right)$	Adjacent Categories
Rasch Partial Credit Model (PCM) (Masters, 1982)	$\theta - \delta_{jk}$	Adjacent Categories
Generalized Partial Credit Model (Muraki, 1992)	$\alpha_j\left(\theta - \delta_{jk}\right)$	Adjacent Categories
Mixed Rating Scale Model (von Davier & Rost, 1995)	$k\theta_c - k\delta_{jc} - \tau_{kc}$, with $\tau_{kc} = \sum_{s=1}^{k} \tau_{sc}$, where τ_{sc} is the cumulated step parameter (one set of steps is assumed to apply to all items)	Adjacent Categories
Mixed Partial Credit Model) (Rost, 1991)	$k\theta_c - \delta_{jkc}$, with $\delta_{jkc} = \sum_{s=1}^{k} \tau_{jsc}$, where τ_{jsc} is the cumulated step parameter for item *j*	Adjacent Categories
Equidistant Rasch Model (Andrich, 1982)	$\theta - \beta_j + \lambda_j\left(m - k\right)$, where λ_j is half the distance between two adjacent steps for item *j*	Adjacent Categories

(table continues)

TABLE I (CONTINUED)

SELECTED SYSTEMATIC COMPONENTS FOR *H* TO REPRESENT DIFFERENT IRT MODELS

Model	Systematic Component (η)	$P\left(Y_j = k \mid \theta;\boldsymbol{\beta}\right)$
Successive Intervals Model (Rost, 1988)	$\theta - \left(\beta_j + \varsigma_j\left(m-k\right) + \kappa_k\right)$, where ς_j is the degree to which the category step distance for item *i* deviates from the overall step distance across all items, and κ_k is the sum of all category steps assicuated with category 1 to *k*.	Adjacent Categories
Monotone Homogeneity Model (MHM) (Mokken & Lewis, 1982)	θ, where $H_{jk}\left(\theta\right)$ is nondecreasing in each coordinate of θ, for all items j=1,…,J	$H\left(\eta_{jk}\right) - H\left(\eta_{j(k+1)}\right)$
Double Monotonicity Model (Mokken & Lewis, 1982)	θ, where $H_{jk}\left(\theta\right)$ is nondecreasing in each coordinate of θ, $H_{jk}\left(\theta\right)$ and $H_{j(k+1)}\left(\theta\right)$ do not intersect, for all k=1,…,m_j, for each item j=1,…,J.	$H\left(\eta_{jk}\right) - H\left(\eta_{j(k+1)}\right)$
Generalized Linear Mixed Models—IRT (e.g., De Boeck & Wilson, 2004)	$\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\upsilon}_i$	Can be either

Note. θ is latent trait level. For the M-RSM and M-PCM, the latent trait level is assumed to be discrete. All other models assume latent trait level is continuous. α_j is the item discrimination parameter. For RSM and M-GRM, β_j is the overall item difficulty, τ_k is the relative step location for category *k*, and all items share a common set of steps. For PCM and GRM, β_{jk} represents the step location for item *j*'s category *k*, although β_{jk} can be expanded to β_j - τ_{jk}, where β_j is the overall item difficulty, and τ_{jk} is the relative step location for item *j*'s category *k*. For cumulative probability models, it is assumed here that $H\left(\eta_{jk}\right) = P\left(Y_j \geq k \mid \theta;\boldsymbol{\beta}\right)$.

Although the GRM and M-GRM are common parametric IRT models, one example of a nonparametric model, the monotone homogeneity model, falls within this class (Mokken & Lewis, 1982): $\eta_{jk} = \theta$, with the constraint that $H_{jk}(\theta)$ is nondecreasing in each coordinate of θ for all categories $k = 0,1,\ldots,m_j$, and test items $j = 1,\ldots,J$, and, in the case of the double monotonicity model, there is also the assumption that $H_{jk}(\theta)$ and $H_{j(k+1)}(\theta)$ do not intersect for all values θ (i.e., satisfying invariant item ordering).

Adjacent category probability models. One type of adjacent category probability IRT model is the Rasch partial credit model (Masters, 1982):

$$\eta_{jk} = \theta - \delta_{jk},$$

given person latent trait θ, and δ_{jk}, denotes item j's step parameter for category k, which is sometimes decomposed in terms of overall item difficulty and relative category step (i.e., $\delta_{jk} = \beta_j + \tau_{jk}$).

A special case of the PCM is the Rasch rating scale model (Andrich, 1978), which assumes all items share a common set of category steps, with $\delta_{jk} = \beta_j + \tau_k$. Thus, similar to the M-GRM, δ_{jk} decomposes into an overall difficulty for item j and a relative step. The RSM and PCM with relative-step parameterization require a constraint to be placed on the τ_ks for model identification. One such constraint is to require the step estimates to sum to zero. Another type of constraint is to assign a proper prior on the distribution of the step estimates, such as $\tau_k \underset{iid}{\sim} N(0,\sigma_\tau^2)$.

GLMM perspective. Most, if not all, current IRT models can be considered members of the general class of mixed generalized linear models (e.g., McCullagh & Nelder, 1989); specifically, the subclass of mixed ordinal regression models, with inverse link $H(\eta): \mathbb{R} \to (0,1)$, systematic component η, and link $h(\cdot) = H^{-1}(\cdot)$ (e.g., De Boeck & Wilson, 2004). Previously, we considered specific forms for the systematic component, but the class of generalized linear mixed ordinal models assumes the general systematic component:

$$\boldsymbol{\eta}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{Z}_t\boldsymbol{\upsilon}_t,$$

where $\boldsymbol{\eta}_t = (\eta_1, \eta_2, ..., \eta_N)'$, for groups $t = 1, \ldots, N$ (where group can be a person), $\mathbf{X} = (x_{tj})_{N \times p}$ is the fixed-effect design matrix corresponding to the column vector of p regression coefficients, $\boldsymbol{\beta}$, $\mathbf{Z} = (z_{tj})_{N \times q}$ is the random-effect design matrix corresponding to the column vector of q random effects, $\boldsymbol{\upsilon}_1, ..., \boldsymbol{\upsilon}_t, \ldots, \boldsymbol{\upsilon}_N \underset{iid}{\sim} G(\boldsymbol{\upsilon})$ for some assumed distribution G, typically a multivariate normal distribution. Also, a nonparametric (e.g., DP) model can be assumed for G (Kleinman & Ibrahim, 1998a, 1998b).

From the perspective of an IRT model, the fixed-effect design matrix **X** contains information pertaining to item covariates (e.g., item difficulty/easiness level and/or category step indicators); person covariates (e.g., gender, ethnicity, and/or socioeconomic status); and/or item-by-person characteristic interactions. **Z** may also contain information about item, person, and item-by-person characteristic interactions. In a traditional IRT model within GLMM, **X** contains information pertaining to item (dummy) indicators and category indicators, and only one random effect is specified, which represents the latent trait, with **Z** a column vector of 1s. When item indicators are included in **Z**, item difficulty and/or step parameter estimates are considered to be random rather than fixed, constituting the random-item IRT model (De Boeck, 2008).

Parameter Estimation

For simplicity, but with no loss of generality, to estimate the population parameter of an IRT model based on available item-responses data, and to describe the different estimation methods, the data are notated as $\mathbf{Y} = (y_{tj})_{N\times J}$, a matrix of responses to J items from a sample of N persons from a population. Across the various estimation methods discussed in the sections that follow, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_t, \ldots, \theta_N)$ represent a column vector or latent trait and $\boldsymbol{\beta}$ represent a column vector of all other parameters in the IRT model (e.g., item, category parameters, etc.). Note that $\boldsymbol{\beta}$ could contain parameters that describe the latent trait distribution (e.g., the mean and variance, μ_θ and τ_θ, respectively, to model $G(\theta) \equiv \mathrm{N}(\theta \mid \mu_\theta, \tau_\theta)$, or more generally, the infinite-dimensional parameter, G). In addition, $(\boldsymbol{\theta}, \boldsymbol{\beta})$ represents all parameters in the model, and the set of all possible values of $(\boldsymbol{\theta}, \boldsymbol{\beta})$ is denoted by Ω.

(Joint) Maximum Likelihood Estimation

The maximum likelihood estimate of $(\boldsymbol{\theta}, \boldsymbol{\beta})$, denoted by $(\widehat{\boldsymbol{\theta}, \boldsymbol{\beta}})$ is:

$$(\widehat{\boldsymbol{\theta}, \boldsymbol{\beta}}) = \underset{(\boldsymbol{\theta}, \boldsymbol{\beta})\in\Omega}{\arg\max} \prod_{t=1}^{N}\prod_{j}^{J} P_j(y_{tj} \mid \theta_t, \boldsymbol{\beta}),$$

and the inverse of the second derivative matrix of $\log\prod_{t=1}^{N}\prod_{j=1}^{J} P_j(y_{tj} \mid \theta_t, \beta)$ evaluated at $(\theta_t, \beta) = (\hat{\theta}_t, \hat{\beta})$ gives the asymptotic sampling variances for $(\widehat{\boldsymbol{\theta}, \boldsymbol{\beta}})$.

Conditional Maximum Likelihood Estimation

Conditional maximum likelihood estimation (CMLE) allows for the estimation of $\boldsymbol{\beta}$ after the latent trait parameters are conditioned out using sufficient statistics. Because sufficient statistics are required, this approach apparently is appropriate only for Rasch models, in which

the person's total score (i.e., the sum of a person's item-level scores), Y_{+}, is a sufficient statistic for latent trait, provided the data fit the model. The CMLE of $\boldsymbol{\beta}$ is obtained via:

$$\widehat{\beta} = \arg\max_{\boldsymbol{\beta}\in\Omega} \prod_{t=1}^{N}\prod_{j=1}^{J} P\left(y_{tj} \mid y_{+t}, \boldsymbol{\beta}\right),$$

where $y_{+t} = \sum_{j}^{J} y_{tj}$ represents the total score for person t. Taking the inverse of the second derivative of $\log\prod_{t=1}^{N}\prod_{j=1}^{J} P\left(y_{tj} \mid Y_{+t}, \widehat{\boldsymbol{\beta}}\right)$ yields the asymptotic variances of $\widehat{\boldsymbol{\beta}}$.

Marginal Maximum Likelihood Estimation

The marginal maximum likelihood estimate is:

$$\widehat{\boldsymbol{\beta}} = \arg\max \prod_{t=1}^{N}\left\{\int \prod_{j=1}^{J} P_j\left(y_{tj} \mid \theta_t, \boldsymbol{\beta}\right) dG\left(\theta_t\right)\right\}, \tag{1.3}$$

and the inverse of the second derivative of Equation 1.3 yields the asymptotic sampling variances of $\widehat{\boldsymbol{\beta}}$. The latent trait parameters $\hat{\boldsymbol{\theta}}$ are estimated in a second stage, given $\widehat{\boldsymbol{\beta}}$, through maximum likelihood, maximum a-posteriori, or expected a-posteriori scoring (see Embretson & Reise, 2000).

Bayesian Inference

Parameter inferences from a Bayesian perspective proceed by finding a solution to Bayes's theorem, which is given by:

$$P\left(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{Y}\right) = \frac{\left\{\prod_{t=1}^{N}\prod_{j=1}^{J} P_j\left(y_{tj} \mid \theta_t, \boldsymbol{\beta}\right)\right\} P\left(\boldsymbol{\theta}, \boldsymbol{\beta}\right)}{\int\left\{\prod_{t=1}^{N}\prod_{j=1}^{J} P_j\left(y_{tj} \mid \theta_t, \boldsymbol{\beta}\right)\right\} P\left(\boldsymbol{\theta}, \boldsymbol{\beta}\right) \mathrm{d}\left(\boldsymbol{\theta}, \boldsymbol{\beta}\right)},$$

where $P(\boldsymbol{\theta},\boldsymbol{\beta})$ is the prior probability density and represents a prior belief about the true values of $(\boldsymbol{\theta},\boldsymbol{\beta})$ underlying the data. Markov Chain Monte Carlo methods, such as Gibbs sampling, Metropolis-Hastings sampling, or slice sampling, are used to simulate from $P(\boldsymbol{\theta},\boldsymbol{\beta} \mid \mathbf{Y})$ to perform inferences on $(\boldsymbol{\theta},\boldsymbol{\beta})$ (Robert & Casella, 2004). One key difference between Bayesian inference and the previously noted maximum likelihood estimation lies in the assumption about the parameters. That is, in Bayesian inference, the parameters are treated as random while within a frequentist approach, the parameters are treated as fixed. By treating the parameters as random, issues of model identification can be avoided with complex models, such as infinite mixture models, and the uncertainty in the parameters are taken into account in the probability model. These issues and Bayesian inference in general are explored in more detail in the following chapter.

Formal Definition of DIF and DSF

Recall that the response to item j is a random variable, Y_j, with a sample space $\{k = 0, 1,\ldots, m_j\}$ where $(m_j \geq 1)$, and is a manifestation of a latent variable, θ, often referred to as the latent trait or ability level. An item is said not to have DIF if the cumulative distribution for $Y_j = k$, given the latent trait, θ, and other person characteristics that are represented by a discrete or continuous random variable W, is the same as the cumulative distribution for $Y_j = k$, given only θ; that is:

$$F\left(Y_j = k \mid \theta, W = w\right) = F\left(Y_j = k \mid \theta\right), \text{ for all } k, \theta, \text{ and } w. \tag{1.4}$$

Otherwise, the item is said to have DIF (Penfield & Camilli, 2007).

To test for item invariance at the category level, a researcher can determine whether Equation 1.4 is true by examining the $K - 1$ category step parameters that govern the ISRF for invariance across W (Camilli & Shepard, 1994; Hambleton & Swaminathan, 1985; Lord, 1980), a process referred to as DSF analysis (Penfield, 2006, 2007) when $K \geq 3$. For rating data, one or all $K - 1$ category step parameters may lack invariance, and the effect of W (i.e., magnitude and direction) on step parameters lacking invariance may or may not differ across them. Thus, I refer to an item as containing overall-level DIF when the effect of W is the same on all $K - 1$ step parameters for an item. Also, I refer to an item as containing DSF when only a subset of $K - 1$ step parameters for an item lacks invariance across W or when all $K - 1$ step parameters lack invariance but the effect of W varies across them.

Parameters Tested for Overall-Level DIF and DSF

As noted previously, in rating scale data, item j can be characterized by a discrimination, α_j, and $K - 1$ step-difficulty parameters, δ_{jk}, or an overall item difficulty level and relative category step (i.e., $\delta_{jk} = \beta_j + \tau_{jk}$). To examine an item for overall DIF, the overall item difficulty level parameter, β_j, is examined for invariance across W. Put more simply, let W represent two subgroups. Item j is said to have overall-level DIF when the difference in the overall item difficulty level estimates for the two groups is statistically significant (i.e., $\beta_{j(W = Group\ 1)} \neq \beta_{j(W = Group\ 2)}$), with $\beta_{j(W = Group\ 1)} > \beta_{j(W = Group\ 2)}$, indicating that the item overall is more difficult for Group 1 compared with Group 2, and vice versa, when $\beta_{j(W = Group\ 1)} < \beta_{j(W = Group\ 2)}$. For IRT models that allow items to vary in terms of discrimination (e.g., GRM and M-GRM), overall-level DIF in item j can also result from a statistically significant difference in the discrimination parameter estimates, α_j, between the two groups (i.e., $\alpha_{j(W = Group\ 1)} \neq \alpha_{j(W = Group\ 2)}$), with a steeper

slope for the ISRFs for Group 1 when $\alpha_{j(W = Group\ 1)} > \alpha_{j(W = Group\ 2)}$, and a less steep slope when $\alpha_{j(W = Group\ 1)} < \alpha_{j(W = Group\ 2)}$.

To detect DSF, each of the $K - 1$ category step parameter estimates are individually examined for invariance across W rather than tested as a set. Again, assuming that W represents two groups, the k^{th} category step parameter is said to have DSF when the difference in the k^{th} step estimates, δ_{jk}, for the two groups is statistically significant (i.e., $\delta_{jk(W = Group\ 1)} \neq \delta_{jk(W = Group\ 2)}$), with $\delta_{jk(W = Group\ 1)} > \delta_{jk(W = Group\ 2)}$, indicating that endorsing category k or higher is more difficult for Group 1 compared with Group 2, and vice versa when $\delta_{jk(W = Group\ 1)} < \delta_{jk(W = Group\ 2)}$. When all items share a common set of category steps, the k^{th} step parameter has DSF for all items when $\tau_{k(W = Group\ 1)} \neq \tau_{k(W = Group\ 2)}$.

IRT Models that Can Detect DSF and Produce Adjusted Scores

Many statistical methods have been employed to assess overall-level DIF for polytomous items (e.g., Chang, Mazzeo, & Roussos, 1996; Dorans & Schmitt, 1993; Flowers, Oshima, & Raju, 1999; Hedeker, Berbaum, & Mermelstein, 2006; Kim & Cohen, 1998; Liu & Agresti, 1996; Rossi, Gilula, & Allenby, 2001; Somes, 1986; Wang, 2004; Williams & Beretvas, 2006). The drawback of overall-level DIF statistics is that they have a relatively low power to detect DIF items when the effect of W differs in direction and magnitude across categories (Ankenmann, Witt, & Dunbar, 1999; Chang, et al., 1996; Penfield & Algina, 2003; Wang & Su, 2004). Moreover, when an overall-level DIF statistic does identifying an item as having DIF, it does not indicate which categories may be contributing to the DIF if all categories do not equally contribute to the DIF (Penfield, Alvarez, & Lee, 2009; Penfield, Gattamorta, & Childs, 2009). For these reasons, I focus on reviewing the literature that pertains to IRT models used in DSF analysis, because DSF analysis reveals which category or categories may be problematic, and

has been demonstrated through simulation studies to possess more power to detect problematic items (Penfield, 2007). For those interested in information on overall-level DIF statistical methods for rating data, the aforementioned works are recommended.

When DSF analysis is conducted, one often relies on a statistical test or global model fit measure. Thus, in this section, I first review the statistical significance tests and global model fit measures used in DSF analysis. I then review traditional IRT models and GLMM-IRT models used to conduct DSF analysis.

Statistical Tests of Parameters

Mahalanobis distance statistic, D^2. The Mahalanobis distance statistic (Cohen, Kim, & Baker, 1993), also referred to as the Wald test, simultaneously tests whether a set of step parameters lacks invariance across *W* at a statistically significant level. The D^2 is based on the vector of the differences in the parameters estimated for two groups, $\hat{\mathbf{v}}$, and its corresponding variance-covariance matrix, $\hat{\Sigma}$ (i.e., $D^2 = \hat{\mathbf{v}}'\hat{\Sigma}^{-1}\hat{\mathbf{v}}$). D^2 tests the null hypothesis of $\mathbf{v}=\left(0\right)_{s\times 1}$, where *s* is the number of parameters tested. D^2 is distributed asymptotically as a chi-square with *s* degrees of freedom. If D^2 is statistically significant at a specified significance level, it reveals only that at least one of the step estimates differs between two groups.

Standardized difference. The standardized difference (*z*) tests whether the difference in the k^{th} step estimates for item *j* between two groups is statistically significant at a specified significance level. This is achieved by taking the difference between the estimates and dividing it by the pooled standard error; that is,

$$z = \frac{\left(\hat{\delta}_{jk(w=group\ 1)} - \hat{\delta}_{jk(w=group\ 2)}\right)}{\sqrt{\mathrm{var}\left(\delta_{jk(w=group\ 1)}\right) + \mathrm{var}\left(\delta_{jk(w=group\ 2)}\right)}}$$

(Cohen, et al., 1993; Lord, 1980; Wright & Masters, 1982). The *z* test tests the null hypothesis of $H_0 : \delta_{jk(w=group\,1)} - \delta_{jk(w=group\,2)} = 0$. The test statistic is then compared with the standard normal distribution, and if *z* is statistically significant, then the k^{th} step for item *j* is said to have DSF.

Measures of Model-Fit Comparisons

The likelihood ratio test (LRT) makes it possible to compare the fit of two nested models and indirectly test the effect of *W* on a single category step or a set of step estimates for statistical significance (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). When conducting an LRT, two models are fit to the data and then compared. Let $(\boldsymbol{\theta},\boldsymbol{\beta})_1$ represent a vector of p_1 parameters in the augmented model, which allows step parameter(s) under examination for lack of invariance to vary as a function of *W*, and $(\boldsymbol{\theta},\boldsymbol{\beta})_2$ represent a vector of p_2 parameters in the restricted model, which constrains some or all of the estimates for the parameters tested to be the same across all *W*, enabling $(\boldsymbol{\theta},\boldsymbol{\beta})_2$ to be contained in $(\boldsymbol{\theta},\boldsymbol{\beta})_1$. Based on the deviance,

$$D(\boldsymbol{\theta},\boldsymbol{\beta}) = -2\sum_{t=1}^{N}\sum_{j=1}^{J}\log P\left[y_{tj}\,|\,(\boldsymbol{\theta},\boldsymbol{\beta})\right],$$

and the LRT statistic is $LRT_{12} = D\widehat{(\boldsymbol{\theta},\boldsymbol{\beta})}_2 - D\widehat{(\boldsymbol{\theta},\boldsymbol{\beta})}_1$, where $D\widehat{(\boldsymbol{\theta},\boldsymbol{\beta})}_1$ and $D\widehat{(\boldsymbol{\theta},\boldsymbol{\beta})}_2$ represent the point-estimates for their respective parameters, and $LRT_{12} \geq 0$. LRT_{12} tests the null hypothesis that the target items are free of DIF/DSF, with the test statistic asymptotically following a chi-square distribution with degrees of freedom equal to the difference in the estimated number of parameters between the larger model and the nested smaller model.

If the null hypothesis is rejected, it can be concluded that the restricted model does not fit the data as well as the augmented model and that at least one step parameter has DSF if multiple-

step parameters are tested. Otherwise, the restricted model fits the data as well as the augmented model and DSF is not present in the examined step parameters. It is important to note that, even for large samples, the LRT tends to be biased in favor of the model with more parameters (Gelfand & Dey, 1994).

Step parameters can also be examined for invariance by comparing the model fit measures (MFM) between two models, such as an augmented model and a restricted model, although the measures I present in this section do not require $\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)_2$ to be contained in $\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)_1$.

The Akaike information criteria (AIC) (Akaike, 1974) is calculated by:

$$AIC = D\widehat{\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)} + 2p\,,$$

where p is the number of parameters estimated in the model. The Bayesian information criteria (BIC) (Schwarz, 1978) is calculated by:

$$BIC = -2B_{01} = D\widehat{\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)} + \log\left(n\right)p\,,$$

where B_{01} is the Bayes factor.

Within a Bayesian inference approach, the deviance information criterion (Spiegelhalter, Best, Carlin, & van der Linde, 2002) is calculated by:

$$DIC = D\overline{\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)} + 2\left\{\overline{D\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)} - D\overline{\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)}\right\},$$

where $D\overline{\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)}$ is the deviance conditioned on the posterior mean estimate $\overline{\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)}$, and $\overline{D\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)}$ is the average of the deviance taken with respect to the posterior distribution, $\pi\left(\boldsymbol{\theta},\boldsymbol{\beta}\,|\,\mathbf{Y}\right)$.

The leave-one-out cross validation (LOOCV) (Geisser & Eddy, 1979) approach in a Bayesian step is calculated through:

$$LOOCV = \sum_{t=1}^{T}\sum_{j=1}^{J}\int \log \mathrm{P}\left[y_{tj} \mid \theta, \boldsymbol{\beta} \right] d\Pi\left(\theta, \boldsymbol{\beta} \mid \mathbf{Y}_{(-tj)}\right),$$

where $\Pi\left(\theta, \boldsymbol{\beta} \mid \mathbf{Y}_{(-tj)}\right)$ is a measure of the posterior distribution conditional on all data except the j^{th} response for person t. The LOOCV under a Bayesian step is also referred to as the log predicted marginal likelihood (LPML). For a model under a non-Bayesian step, the leave-one-out log predictive likelihood is calculated through:

$$LOOCV = \sum_{t=1}^{N}\sum_{j=1}^{J}\int \log \mathrm{P}\left[y_{tj} \mid \hat{\theta}_{(-tj)}, \hat{\boldsymbol{\beta}}_{(-tj)} \right] \mathrm{d}y_{tj},$$

where $\hat{\theta}_{(-tj)}, \hat{\boldsymbol{\beta}}_{(-tj)}$ are the estimates for their respective parameters based on data with the j^{th} response for person t removed from the analysis. LPML and LOOCV are on a common log-likelihood scale, which makes it possible to directly compare the predictive utility between a model under point estimation and a model under full Bayesian estimation.

The $D(m)$ criterion (Gelfand & Ghosh, 1998), which is a mean-squared predictive criterion, is another predictive performance index that allows for the comparison of models under point estimation and full Bayesian estimation. Given model $\underline{m}$, where the parameters are from Bayesian estimation, the criterion is defined by:

$$\begin{aligned} D\left(\underline{m}\right) &= \sum_{p=1}^{N}\sum_{j=1}^{J}\left[y_{pj} - \mathrm{E}\left(Y_{pj} \mid \mathbf{x}, \underline{m}\right)\right]^2 + \sum_{p=1}^{N}\sum_{j=1}^{J}\mathrm{Var}\left(Y_{pj} \mid \mathbf{x}, \underline{m}\right) \\ &= \mathrm{GF}\left(\underline{m}\right) + \mathrm{Pen}\left(\underline{m}\right), \end{aligned}$$

where y_{pj} is person p's observed response for item j, $\mathrm{E}(Y_{pj}|\underline{m})$ is the expected value y_{pj} given model $\underline{m}$, and $\mathrm{Var}(Y_{pj}|\underline{m})$ is the variance for Y_{pj} given $\underline{m}$. The $\mathrm{GF}(\underline{m})$ indicates the goodness-of-fit to the sample data, $\mathcal{D}_n$, at hand. The term, $\mathrm{Pen}(\underline{m})$, indicates the penalty and is large when the

model is over- or under-fitting the sample data. Lower $D(\underline{m})$ values indicate better fitting models to the data.

For a non-Bayesian model having point estimate $\hat{\varphi}_n = \hat{\varphi}\left(Data\right)$, such as a maximum-likelihood estimate, the $D(\underline{m})$ criterion is estimated via $\hat{\mathrm{E}}\left(Y_{pj} \mid \mathbf{x}_{pj}, \underline{m}\right) = \mathrm{E}\left(Y_{pj} \mid \mathbf{x}_{pj}, \underline{m}, \hat{\varphi}\right)$ and $\widehat{\mathrm{Var}}\left(Y_{pj} \mid \mathbf{x}_{pj}, \underline{m}\right) = \mathrm{Var}\left(Y_{pj} \mid \mathbf{x}_{pj}, \underline{m}, \hat{\varphi}\right)$ (Gelfand & Ghosh, 1998).

Whenever $MFM_1 < MFM_2$ for two alternative models, wherein *MFM* is one of the aforementioned model-fit measures, the augmented model is preferred to the restricted model, which suggests that at least one of the examined step parameters has DSF; otherwise, the restricted model is preferred to the augmented, which suggests that none of the examined step parameters have DSF.

The BIC is asymptotically consistent (Kuha, 2004; Schwarz, 1978); that is, the model-fit measure leads one to choose the true model as sample size increases to ∞; although for large sample sizes, the BIC tends to select the simpler model because the penalty takes the sample size into consideration (Kang, Cohen, & Sung, 2009; Kuha, 2004). The AIC, on the other hand, is not asymptotically consistent (Schwarz, 1978; Sclove, 1987), and tends to be biased in favor of the more complex model because the penalty does not take the sample size into consideration (Janssen & De Boeck, 1999; Kuha, 2004). The LOOCV has been shown to be asymptotically equivalent to the AIC (Stone, 1977). Thus, the limitations of AIC apply to LOOCV as well. In terms of asymptotic consistency and model selection, it is possible to adopt two viewpoints. One viewpoint is that the measure of model fit should be asymptotically consistent, a fact that would preclude the use of the AIC in model selection. The other viewpoint is that, for a finite sample,

the concern should be predictive validity. The concern for asymptotic consistency may not be as critical in the latter viewpoint as it is in the former.

Traditional IRT for DSF Analysis

Multiple-Group IRT

In multiple-group IRT (Lord, 1980; Wright & Masters, 1982), W is assumed to be discrete (e.g., representing gender or ethnicity). This model makes it possible to simultaneously estimate category step parameters for each group of interest and place all parameter estimates on a common metric, an approach that allows for direct comparisons of step parameter estimates across groups (Embretson & Reise, 2000). Multiple-group IRT allows for the groups' latent trait distribution in the population to vary, which makes possible a more accurate detection of any differences in item parameter estimates (Thissen, et al., 1986). That is, $G(\theta)$ becomes $G_w(\theta)$, with a typical assumption about the form of the distributions in the population, that each group has a normal distribution with some mean and variance, with the means and variances allowed to vary across the groups.

Multiple-group IRT requires the use of a group-linking procedure to establish a common metric for the parameter estimates across the groups. In the sections that follow, I present two common group-linking procedures.

Equal-mean-difficulty procedure. In the equal-mean-difficulty (EMD) approach, the mean of the item-difficulty parameter estimates is constrained so that it is the same across the groups (Muraki, 1999; Wang, 2004, 2007). To conduct a DSF analysis with this procedure, for each group, one would estimate J sets of category steps if each item is assumed to possess its own set of steps, and estimate only one set of category steps if a single set is assumed to apply to all items, while simultaneously constraining the mean of the overall item difficulty estimates so

that it is equal across all groups. Then, with a statistical test, one can compare the step estimates between two groups or rely on a model-fit measure comparison to determine whether DSF is present.

In a simulation study (Wang, 2004), the EMD procedure led to an increased Type I error rate in detection of DSF when the rating data were generated without equality of mean item difficulty between the two groups. The Type I error rate increased as the true mean item difficulty difference between the two groups increased. Moreover, the magnitudes of the DSF effect were overestimated. The advantage of the EMD procedure is that it is not necessary to identify an item or a set of items that are invariant across the groups (constituting an approach described in the section that follows). The disadvantage of the EMD method is that it is appropriate under only three conditions: (a) when all items are invariant across the groups; (b) when one category step estimate favors one group, another category step estimate must favor the other group with the same magnitude; and (c) all groups must be presented with the same set of items (Wang, 2004, 2007).

Anchoring procedure. In the anchoring procedure, the item parameter estimates for a subset of items are constrained so that they are the same across all groups, whereas the parameter estimates for the remaining items are freely estimated for each group. The subset of items is considered to be the anchor items (Thissen, et al., 1988, 1993; Wang, 2004, 2007). The number of anchor items can range from 1 to $J - 1$. To test for DSF, one can either conduct a statistical significance test or perform model comparisons.

Simulation studies that analyzed data generated with multiple-group GRM and PCM (Wang, 2004; Wang & Yeh, 2003) showed that when $J - 1$ items served as anchors, the Type I error rate was higher when some of the DSF-containing items were included as anchor items, and

the error rate was lower when fewer than $J - 1$ items served as anchors but all were invariant across groups. In addition, the magnitudes of the DSF effects were overestimated when the items used as anchors lacked invariance across the groups. When a single invariant item served as the anchor, the Type I error rate remained near the nominal level of .05, and the number of items ranged from 25 to 30. As the number of anchor items increased to 10 with all anchor items invariant between the two groups, the Type I error rate remained the same while the power to identify problematic steps increased; however, with respect to power, a point of diminishing returns was observed after four items.

In another simulation study (Kim & Cohen, 1998), when $J - 1$ invariant items served as anchors, the Type I error rate in DSF detection remained near the nominal level of .05 across different conditions in which the sample size and latent trait distributions varied between two groups.

Conclusions

An advantage of the multiple-group IRT is that adjusted latent trait estimates can be produced when DSF is present. The disadvantages are that (a) the multiple-group approach assumes that the grouping variable explains all of the DSF if DSF is present, (b) group information must be available, (c) W cannot be a continuous person characteristic, (d) interactions among grouping variables cannot be explored, and (e) all items cannot be simultaneously tested when a subset of items serve as anchors.

Mixture Item Response Models

Mixture IRT assumes that at least two unobserved subpopulations exist in the population from which the sample was selected, with the possibility that their latent trait levels have

different distributions. The overall item and/or step parameters might or might not differ across these latent subpopulations, and latent trait levels are discrete (von Davier & Yamamoto, 2004). Some models that fall under this category include Yamamoto's HYBRID model (Yamamoto, 1987), the mixed Rasch model (Rost, 1990; von Davier & Rost, 1995), and the discrete generalized partial credit model (von Davier & Yamamoto, 2004). To examine for DSF, a model in which a single latent class is assumed to exist (i.e., the restricted model) is compared with a model in which more than one latent class is assumed to exist (i.e. the augmented model), with the step parameter estimates allowed to vary across the latent classes. The measures of model fit between the two models are then compared.

Rost, Cartensen, and von Davier (1997) analyzed rating data with the mixed PCM, revealing that DSF was present in every item across two latent classes. Wagner-Menghin (2007) analyzed a data set with the mixed RSM, and revealed that the step parameter estimates were disordered for one latent class, but were monotonically ordered as the rating categories increased for the other latent class.

The advantages of mixture IRT are that (a) group information does not need to be known or may be known for only some of the persons (von Davier & Yamamoto, 2004), (b) if the number of possible latent groups is not known, the analysis can estimate the number of possible latent classes based on the data (von Davier & Yamamoto, 2004), and (c) latent trait estimates can be produced even when DSF is present.

The disadvantages with the mixture IRT are that (a) when the step estimates differ, the latent classes are assumed to explain all of the DSF present in a step, (b) step estimate invariance cannot be explored across a continuous covariate, and (c) the same number of latent groups are assumed across all items.

Item Response Theory within Generalized Linear Mixed Models for DSF Analysis

Fixed-Effect Approach

To determine whether category step parameters contain DSF, a model fit comparison can be made between an augmented model, in which person characteristics by item step interactions are included in $\boldsymbol{X}$, and a restricted model, in which the interaction is not included in $\boldsymbol{X}$.

In a simulation study in which data were generated for 15 items, with one item assumed to contain DSF steps, the power to detect DSF with the GLMM-GRM ranged from .95 to 1.00 regardless of sample size. The familywise Type I error ranged from .02 to .15 when $N = 500$ and from >.001 to .01 when $N = 1{,}250$ (Vaughn, 2006).

Random-Step Parameter Approach

GLMM-IRT allows the category step parameters to be treated as random (Johnson, 2003; Tutz & Hennevogl, 1996), and a DSF analysis can be conducted in this manner. In an augmented model, the category step indicators that correspond to the steps tested for DSF are included in $\boldsymbol{Z}$. When the variance component associated with δ_{jk} is zero, the step estimates do not vary across the persons, a fact that suggests that DSF is not present in item j's k^{th} step. This is because when a step parameter is invariant across all persons, it will also be invariant across any W.

In a simulation study, Johnson (2003) showed that the model that treated the step parameters as fixed effects (i.e., a fixed-step model) and the model that treated the step parameters as random (i.e., a random-step model) yielded similar estimates for the overall intercept and the variance for the random intercept when the data were generated under a condition in which the step estimates did not vary across the persons. However, when the data were generated under the condition in which the step parameter estimates varied across the persons, the fixed-step model underestimated the overall intercept estimate and the variance of

the random intercept component. In addition, the magnitude of the underestimation increased as heterogeneity in step estimates across persons increased. Moreover, the random-step model led to unbiased results and better global model fit measures than did the fixed-step model.

When Tutz and Hennevogl (1996) analyzed two real data sets with the random step model, they found statistically significant variation in the step estimates, which suggested that not all persons were using the rating categories in a similar manner. In addition, the authors showed that the magnitudes for the fixed covariates increased when the step parameters were treated as random, a finding that is in line within the theoretical inferences of the field from which the data were obtained.

Random-Group GLMM-IRT

One way to establish a common metric across groups is to use either the EMD method or the anchoring method. The EMD method is appropriate only under very limited conditions. The anchoring method requires finding four items on which to anchor, which can be challenging, especially as the number of groups increases, because as noted previously, all step parameters for each item must be invariant across all groups (Stark, Chernyshenko, & Drasgow, 2006). By specifying a three-level IRT model in which the item responses are nested within persons, and the persons are nested within groups, which are treated as random, the need to anchor the groups on a set of items is circumvented (De Jong, et al., 2007). Instead, a prior assumption is placed on the distribution of the step and discrimination estimates. That is, the k^{th} step and discrimination parameters for item j and group w are assumed to be distributed as:

$$\delta_{jwk} \sim N\left(\delta_{jk}, \sigma_{\sigma}^{2}\right),\ \delta_{jw1} \leq \ldots \delta_{jwk} \leq \ldots \leq \delta_{jwm_j},\ \text{for all}\ j,$$
$$\alpha_{jw} \sim N\left(\alpha_{j}, \sigma_{\alpha}^{2}\right) I\left(\alpha_{jw} > 0\right),$$

where δ_{jk} is the grand mean for item *j*'s k^{th} step, and α_{jm} is the grand mean for item *j*'s discrimination parameter. Each group's step estimate is allowed to deviate from this grand mean. When σ_{σ}^{2} is not statistically significantly different from 0, then DSF is not present in the k^{th} step for item *j*; otherwise the step is said to have DSF. In a simulation study involving 10 groups (De Jong et al., 2007), when rating data were generated under the condition in which all items were noninvariant across the 10 groups, the random group model recovered all item parameters, group latent means and variances, and correctly identified all DSF.

Mixture MGLMM Multidimensional IRT

Although the random group GLMM-IRT overcomes the need to anchor groups on a common set of items, it assumes that all groups belong to a single metagroup, which constitutes a restrictive prior, and if inappropriate, it might obscure any noninvariance in item parameters (De Jong & Steenkamp, 2010). To overcome this restriction, De Jong and Steenkamp (2010) presented a mixture GLMM-multidimensional IRT model for ordinal outcomes. Each group, *w*, is assumed to belong to one of *L* latent classes, and a group's step parameters are assumed to be drawn from the heterogeneity distribution associated with the latent class to which the group belongs. If $e_w = l$, then the distribution of the step and discrimination parameters are:

$$\delta_{jwk} \mid e_w = l \sim N\left(\delta_{jk}^{(l)}, \sigma_{\sigma q(k)}^{2,(l)}\right),\ \delta_{jw1} \le \ldots \delta_{jwk} \le \ldots \le \delta_{jwm_j},\ \text{for all } j$$

$$\alpha_{jwq} \mid e_w = l \sim N\left(\alpha_{jd}^{(l)}, \sigma_{\alpha q(k)}^{2,(l)}\right) I\left(\alpha_{jwq} > 0\right),$$

where $q(k)$ is the dimension for which item *j* is an indicator. The mixtures are indicated by $l = 1,\ldots, L$, and e_w is assumed to be drawn from a multinomial distribution, $e_j \sim \text{multinomial}(\boldsymbol{\psi}, 1)$, where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_L)$ and $\sum_{l=1}^{L} \psi_l = 1$.

If a step-parameter estimate varies across groups within a latent class or across latent groups, then that step is said to have DSF. In this model, it is possible that, when multiple latent classes are assumed to be present, the k^{th} step for item j may display DSF across groups within one latent class, but not for another. When only one latent class is present and no statistically significant variation in the estimate for the k^{th} step for item j is observed across the groups, then that step is considered not to have DSF.

In a simulation study, when data were generated under the condition in which two latent classes were present and the construct was assumed to be two-dimensional, the mixture GLMM-multidimensional IRT was a better fit for the data than the random group GLMM-IRT. Moreover, the mixture GLMM-multidimensional IRT model perfectly recovered the latent class membership after a burn-in phase and accurately recovered the item parameters (De Jong & Steenkamp, 2010).

Summary of GLMM-IRT

In addition to the advantages of the traditional IRT models, there are several other potential advantages of the GLMM-IRT. For example, DSF can be examined as a function of known grouping variables, or by modeling latent classes when group association is not known but assumed to exist. After group association or latent classes are accounted for, the amount of variation that remains in step parameter estimates across the persons can be examined, and therefore it is not assumed that group covariates and/or latent class association accounts for all the DSF. The difficulty of finding items on which to anchor the groups is circumvented. W can be continuous, so the DSF investigation is not limited only to categorical variables. Finally, adjusted scores can be produced while controlling for DSF.

Even with the benefits derived from fitting IRT models within GLMM, limitations nevertheless exist. Many of the GLMM-IRT models assume that the true distribution (G) of the random effects follows a parametric distribution. That is, a finite number of parameters describe the true population distribution of the random effects. One problem with assuming a parametric distribution for the random effect is that an incorrect choice of a parametric distribution can lead to poor estimations of the random effects (Verbeke & Lesaffre, 1996) and can possibly lead to erroneous inferences about a person's performance. In addition, if category step parameters are treated as random, inaccurate conclusions could be made about the extent of heterogeneity in step parameter estimates. Another problem with an incorrect parametric distributional choice for the random effects is that slight changes in the chosen parametric distribution for the random effects can lead to fairly large changes in other parameter estimates (Heckman & Singer, 1984). Choosing an incorrect parametric distribution for the random effects, then, can lead to imprecise estimates of the effects of W on step parameters and inaccurate conclusions about the extent of DSF present in an item.

Although mixture GLMM-IRT addresses many of the disadvantages of traditional IRT, it is not without shortcomings. It assumes the same latent class structure for each dimension of a construct that a test is intended to measure, if the item responses for all dimensions are analyzed simultaneously. The problem with this assumption is that different latent structures could exist across dimensions of a latent construct. If true, then inferences about validity concerning scores based on results from a model in which the same latent class structure is presumed across dimensions may be misleading. Moreover, if the correct latent class structure cannot be captured across dimensions, the quality of the adjusted scores could be degraded, because the construct-irrelevant variance present in the measurement process may not be controlled for appropriately.

Approaches That Detect Only DSF

In this section, I present the common odds ratio estimator (CORE) (Penfield, 2007) and the logistic regression approach (French & Miller, 1996) as alternatives when, for some reason, IRT or GLMM-IRT cannot be used to perform a DSF analysis. An example is when the data do not meet the assumptions or the sample size is insufficient for stable parameter estimation. Although these two statistical models fall outside of traditional IRT and GLMM-IRT, they apply dichotomization schemes to rating data so that the rescored data represent step functions analogous to those described in the IRT section. I first present the way in which the data must be rescored to reflect IRT step parameters, and then I review the statistical models.

Rescoring the Data

The responses to an item must be represented through $K - 1$ dummy indicators, and each dummy coding can be performed to be analogous to a cumulative or adjacent category step parameter. For the k^{th} cumulative dummy coding for item j, rescoring to the item responses is performed as follows:

$$Cumulative_{kj} = \begin{cases} 1 \text{ if } y_j \geq k \\ 0 \text{ otherwise} \end{cases}, \quad \text{for } k = 1, 2, \ldots, m_j \text{ when the response option ranges from 0 to } m_j.$$

The second rescoring scheme reflects adjacent category probabilities. For the k^{th} adjacent category dummy coding for item j, rescoring to the item responses is performed as follows:

$$Adjacent_{kj} = \begin{cases} 1 \text{ if } y_j = k \\ 0 \text{ if } y_j = k - 1 \\ \text{otherwise set to missing} \end{cases}, \quad \text{for } k = 1, 2, \ldots, m_j \text{ when the response option ranges from 0 to } m_j.$$

A statistical analysis is then performed on each of the dummy coded indicators.

Common Odds Ratio Estimator

The CORE tests the odds of 1 relative to 0 in $Dummy_{kj}$ (where $Dummy_{kj}$ may be either $cumulative_{kj}$ or $adjacent_{kj}$) across groups. Thus, W must be discrete. To examine item j's k^{th} step parameter for DSF, one must first stratify the total score into R strata to control for performance level and then calculate a conditional odds ratio for two groups at a time, using $Dummy_{kj}$ as the data. The CORE test statistic for the k^{th} step for item j is calculated by:

$$z\left(\hat{\lambda}_{jk}\right) = \frac{\ln\left(\hat{\lambda}_{jk}\right)}{se\left(\hat{\lambda}_{jk}\right)}$$

$$= \ln\left(\frac{\sum_{r=1}^{R} A_{kr} D_{kr} / N_{kr}}{\sum_{r=1}^{R} B_{kr} C_{kr} / N_{kr}}\right) \Bigg/ \sqrt{\frac{\sum_{r=1}^{R} N_r^{-2}\left(A_{kr} D_{kr} + \hat{\alpha}_k B_{kr} C_{kr}\right)\left(A_{kr} + D_{kr} + \hat{\alpha}_k B_{kr} + \hat{\alpha}_k C_{kr}\right)}{2\left(\sum_{r=1}^{R} \frac{A_{kr} D_{kr}}{N_r}\right)^2}},$$

where A_{kr} is the frequency of 1s in $Dummy_{kj}$, B_{kr} is the frequency of 0s in $Dummy_{kj}$ for Group 1 members in stratum r (where $r = 1, \ldots, R$); C_{kr} is the frequency of 1s in $Dummy_{kj}$, and D_{kr} is the frequency of 0s in $Dummy_{kj}$ for Group 2 members in stratum r. The test statistic is asymptotically distributed as a standard normal (Hauck, 1979) and tests the null hypothesis of $H_0 : \lambda_{jk} = 0$, or DSF is not present in item j's k^{th} step.

In a simulation study (Penfield, 2007), the Type I error rate remained near the intended nominal level when the rating data were generated to conform to the GRM under the condition that the latent trait distributions for the two groups were the same. The power to detect DSF when only one of the three steps contained DSF (i.e., $\lambda_{jk} = .6$) ranged from .816 to .957 when the latent trait distributions for the two groups were the same, and .815 to .850 when the two groups' mean of the latent trait distributions differed by one standard deviation. When the data were generated under the condition in which one step had a DSF effect of $\lambda_{jk} = .2$ and the other

step had a DSF effect of $\lambda_{jk} = .4$ (where higher values represent greater DSF), the power to detect the more problematic step ranged from .408 to .650 when the latent trait distributions for the two groups were the same and .351 to .531 when the latent trait distributions differed by one standard deviation. When the data were generated under the GPCM, similar trends in power and Type I error were observed.

An advantage of CORE is that it requires a smaller sample size than IRT and GLMM-IRT (Penfield, 2007; Penfield, Myers, & Wolfe, 2008). Its disadvantages are as follows.

- It loses power when the group latent trait distributions differ with respect to their means.
- The data for item *j* must be analyzed $K - 1$ times, and each analysis assumes independence.
- The impact of missing data has not been addressed with this approach, a critical issue because total score is used to control for performance level.
- The total scores used to represent the performance level are assumed to have been measured without error.
- A DSF analysis can be conducted only across discrete variables.
- This model cannot produce adjusted scores when a step or steps contain DSF.

Logistic Regression Approach

The logistic regression method for detecting DSF at the k^{th} step for item *j* involves modeling the log odds of $Y_{jk} = 1$ relative to $Y_{jk} = 0$ as a function of total score, A, person covariate, W (where W may be a continuous or group dummy indicator), and total score by person covariate interaction, $A \times W$:

$$Log\left(\frac{\Pr\left(Y_{jk}=1\right)}{1-\Pr\left(Y_{jk}=1\right)}\right)=\beta_0+\beta_1 A+\beta_2 W+\beta_3 A*W\ ,$$

where Y_{jk} is $Dummy_{jk}$. β_2 and β_3 inform us whether the location and slope, respectively, for the k^{th} step function varies across W after controlling for performance level.

In a simulation study involving two groups (French & Miller, 1996), in which rating data were generated under the PCM and all steps had the same discrimination level for the two groups, the power to detect DSF in the $n = 500$ condition (with the Type I error rate set at .002) ranged from .77 to .99. When the cumulative and adjacent category rescoring (respectively) was performed on the data, the power ranged from .99 to 1.00, and when $n = 2{,}000$, it was 1.00, regardless of the rescoring scheme.

Although the disadvantages of this statistical approach are similar to those of the CORE, two advantages are that (a) more than two groups can be examined by including additional group dummy indicators, and (b) W may be continuous.

Potential of the Mixture Latent Distribution Model for DSF Analysis

In this section, I review the mixture latent distribution model (Kottas, Müller, & Quintana, 2005), which has not been used specifically for DSF, but has potential for use in modeling and assessing differential category usage. The traditional IRT and GLMM-IRT models heretofore discussed examine whether the step location on the latent trait scale differs between at least two groups while the inverse link function remains fixed (e.g., logistic CDF). Kottas et al. (2005) assigned the category step parameters to fixed values that monotonically increase with the categories while allowing the latent distributions to vary across persons. The univariate version of their model can be represented as:

$$y_i \sim \int_{A(y_i)} N(y_{*i} \mid m_i, s_i)\,\mathrm{d}\,y_{*i},$$

$$(m_1, s_1), \ldots, (m_n, s_n) \mid G \sim_{\text{iid}} G,$$

$$G \sim \mathrm{DP}(\alpha, G_0),$$

Where G is distributed as a Dirichlet Process (DP; Ferguson, 1973) with baseline distribution G_0 being a normal-inverse-gamma distribution, while the A(k) ($k = 0,1,\ldots,m$) are pairwise disjoint subsets whose union is $\Re$. This model accounts for and explains differences in rating category usage without placing any restrictive assumption on the distribution of the latent distribution in the population. Moreover, with respect to predictive accuracy of the ordinal latent dependent variable, this model outperformed an ordinal regression model that parameterized category steps but assumed a fixed inverse link function (Kottas et al., 2005).

Extending the work of Kottas et al. (2005), Karabatsos and Walker (2012) introduced an infinite-mixture Bayesian nonparametric regression model, having covariate dependent mixture weights. For ordinal dependent responses, each kernel of the mixture can be defined by a latent and highly flexible unimodal density, with fixed steps for the ordered categories. Specifically, each unimodal kernel density is assigned a general nonparametric, stick-breaking (e.g., DP) mixture prior (Ishwaran & James, 2002). Karabatsos and Walker showed that, for real and simulated data, their regression model outperformed all other well-known parametric, semiparametric, and nonparametric regression models, in terms of predictive accuracy of the dependent variable, as measured by cross-validated log-likelihood.

Open Problems with Current IRT Models Used in DSF Analysis

IRT models provide a means to examine items for DSF. Moreover, if items are detected to have DSF, those items can be retained and the IRT models can produce adjusted scores. However, the current IRT models used in DSF analyses have noticeable limitations.

Some of the limitations of traditional IRT models include: the challenge of finding anchor items; the person characteristic, W, used to model a category step must be discrete and

known; and assuming W explains all of the DSF if W is a statistically significant predictor of a category step estimate. Although GLMM-IRT models are more flexible than traditional IRT models, they too have notable limitations. The person characteristics must be known when step parameters are treated as fixed effects. Moreover, GLMM-IRT models assume that the true distribution (G) of the random effects is of a known parametric form. That is, a finite number of parameters describe the true population distribution of the random effects. A notable limitation of mixture GLMM-IRT models is that they assume the same latent class structure across all items and for each dimension of a construct that a test is intended to measure.

Hence, to address such issues, I devised a model that identifies clusters of items and persons, and if multiple clusters of persons are present, the items comprising a cluster of items do not have to be the same across all clusters of persons. This is accomplished without the need to find a set of items to serve as anchors. To accomplish this, I define a nonparametric model for the mixing distribution G, which also accounts for covariate-dependent step parameters. Doing so could lead to more accurate detection of noninvariant items at the category level and produce more appropriate adjusted scores.

In Chapter II, I describe my methodology for addressing these issues and introduce a dependent Dirichlet process model for rating data. In Chapter III, I provide a simulation study that evaluates the obtained estimates of the proposed model. In Chapter IV, I describe applications of the proposed model to real rating data. In both Chapters III and IV, I compare the predictive performance of the proposed model against standard IRT models. In Chapter V, I provide additional follow-up analyses to explain the findings in Chapters III and IV. In Chapter VI, I conclude with a discussion of practical implications of the proposed model, as well as its limitations and possible future modeling extensions.

II. METHODS

In Chapter 1, I noted the limitations of traditional item response theory (IRT) models, generalized linear mixed models (GLMM)-IRT, and finite-mixture IRT for use in differential step functioning (DSF) analysis. Specifically, these models traditionally assume that the true distribution of the random effects, *G*, follows a parametric form (i.e., a finite number of parameters describe the shape of the distribution), and when the mixture is discrete, the number of latent classes (clusters) is not covariate dependent. To address the limitations of the traditional mixture GLMM-IRT, in this chapter, I introduce a covariate-dependent infinite-mixture IRT model. The mixing distribution for this model is formed via the multiple Dirichlet process (mDP) (Basu, 2007), which is a type of dependent Dirichlet process (DDP) (MacEachern, 1999, 2001). This model falls within the general framework of the dependent Dirichlet process rating model (DDP-RM) (Fujimoto & Karabatsos, 2014).

An overview of finite-mixture IRT models precedes a discussion of the traditional nonparametric Dirichlet process (DP) model. Then I present the model that is the focus of this thesis, the mDP version of the DDP-RM. I define the parameters and describe the prior distributions for all parameters. I then describe the Markov chain Monte Carlo (MCMC) algorithm that was used to sample the posterior distribution of the model. Lastly, I describe the data sets that I analyzed, as well as the traditional models to which I compared my model in terms of predictive performance, and the criteria I used to judge the performance of my model as it relates to the traditional IRT models.

Finite-Mixture IRT

When performing a DSF analysis with a covariate-independent finite-mixture IRT, the random density function has the general form (e.g., McLachlan & Peel, 2000):

$$f_G\left(y \mid \boldsymbol{\theta}\right) = \int f\left(y \mid \boldsymbol{\theta}, \boldsymbol{\tau}\right) \mathrm{d}G\left(\boldsymbol{\tau}\right).$$

Given that *G* is discrete, the random density could also be expressed as:

$$f_G\left(y \mid \boldsymbol{\theta}\right) = \sum_{h=1}^{H} f\left(y \mid \boldsymbol{\theta}, \boldsymbol{\tau}_h\right) \omega_h, \tag{2.1}$$

where $h = 1,\ldots H$ is the finite number of mixture component indices, *G* the mixing distribution, kernel densities $f\left(y \mid \boldsymbol{\theta}, \boldsymbol{\tau}_h\right)\left(\text{for } h = 1,\ldots,H\right)$, with fixed parameter θ and random parameter $\boldsymbol{\tau}$ that is subject to the mixtures, and mixing weights $\left(\omega_h\right)_{h=1}^{H}$ that sum to 1. As previously noted, a limitation of the finite mixture model is that, for all item category steps (i.e., $\boldsymbol{\tau}$) treated as random, the same number of mixture components (i.e., *H*) is specified. This assumption is appropriate when the data corresponding to all random items indeed reflect DSF across *H* latent classes. However, this assumption is easily violated when some subset of random items are free of DSF and another subset of random items displays DSF between genders. To overcome this limitation, a nonparametric IRT modeling approach based on a DDP is explored. To facilitate the transition to the presented DDP model, I provide an overview of a DP model.

The Dirichlet Process Model

For notational convenience, I denote $\text{normal}\left(\cdot,\cdot\right)$, $\text{normal}_p\left(\cdot,\cdot\right)$, $\text{ig}\left(\cdot,\cdot\right)$, $\text{iw}\left(\cdot,\cdot\right)$, $\text{beta}\left(\cdot,\cdot\right)$ as the probability density functions for the univariate normal, *p*-variate normal, inverse gamma, inverse Wishart, and beta distributions. The inverse gamma distribution is parameterized by shape and rate, and the inverse Wishart distribution is parameterized by degrees of freedom and inverse of a scale matrix. The DP prior provides a means of nonparametrically modeling the form of the random distribution (Ferguson, 1973) *G*, which overcomes one of the limitations (described in Chapter I) of the traditional IRT and GLMM-IRT models. That is, $G \sim DP\left(\alpha, G_0\right)$,

which states that G is distributed as a DP with a precision parameter, $\alpha > 0$, and baseline distribution, G_0, which defines the expectation (mean) of G. A priori, under $DP(\alpha, G_0)$, the marginal distributions of G are Dirichlet distributed as

$$\left(G(A_1),\ldots,G(A_k)\right) \sim \text{Dirichlet}\left(\alpha G_0(A_1),\ldots,\alpha G_0(A_k)\right),$$

for all finite measurable disjoint partitions of $A_1, \ldots A_k$ of Ω; the mean of G is given by $\mathrm{E}\left[G(\cdot)\right] = G_0(\cdot)$; and the variance is obtained through

$$\mathrm{Var}\left[G(\cdot)\right] = \frac{G_0(\cdot)\left[1 - G_0(\cdot)\right]}{\alpha + 1}.$$

Given a set of observed data $\boldsymbol{y}_n = \{y_1,\ldots,y_n\}$, the DP prior distribution gets updated to a posterior distribution, which is also a DP. In the posterior, the marginal distributions of G are Dirichlet distributed as

$$\left[\left(G(A_1),\ldots,G(A_k)\right) \middle| \boldsymbol{y}_n\right] \sim \text{Dirichlet}\left(\alpha G_0(A_1) + n\hat{G}_n(A_1),\ldots,\alpha G_0(A_k) + n\hat{G}_n(A_k)\right),$$

for all finite measurable disjoint partitions of $A_1, \ldots A_k$ of Ω, and where $\hat{G}_n(\cdot)$ is the empirical cumulative distribution function (CDF); the mean of G is given by

$$\mathrm{E}\left[G(\cdot) \middle| \boldsymbol{y}_n\right] = \frac{\alpha G_0(\cdot) + n\hat{F}_n(\cdot)}{\alpha + n} = \bar{G}(\cdot);$$

and the variance is given by

$$\mathrm{Var}\left[G(\cdot) \middle| \boldsymbol{y}_n\right] = \frac{\bar{G}(\cdot)\left[1 - \bar{G}(\cdot)\right]}{\alpha + n}.$$

Stick-breaking Process

Sethuraman (1994) provided a convenient process with which to express the DP prior. This process is referred to as a stick-breaking construction. In the stick-breaking process, the mixture distribution *G* is constructed through

$$G(\cdot) = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h}(\cdot),$$

with stick-breaking weights

$$\omega_h = \upsilon_h \prod_{l=1}^{H-1} (1 - \upsilon_l),$$

for all $h = 1, 2, \ldots,$ summing to 1 with probability 1, with $\upsilon_1, \upsilon_2, \ldots \sim_{i.i.d.} \text{Beta}(1, \alpha)$ and atoms $\theta_1, \theta_2, \ldots \sim_{i.i.d.} G_0$ for $h = 1, 2, \ldots$, while the degenerate distribution $\delta_{\theta_h}(\cdot)$ assigns a probability mass of 1 on the value θ_h. Based on the stick-breaking process representation, it is obvious that the DP supports discrete distributions with probability 1. It should also be noted that a generalized stick-breaking process is applied by drawing a random sequence of values underlying the mixture weights from a general beta distribution in which the shape and scale parameters are allowed to vary; that is, $\upsilon_1, \upsilon_2, \ldots \sim_{i.i.d.} \text{Beta}(a_h, b_h)$. Conceptually, a stick-breaking process can be viewed as starting with a stick of unit length. A piece is broken off this stick and becomes the probability weight for the first mixture component. Of the remaining length of the stick, another piece is broken off and becomes the probability weight for the second mixture component. This process then continues. With each step, the remaining portion of the stick decreases in length, thus leading to a smaller probability of supporting the remaining mixture components. Based on this example, it is easy to see that the stick-breaking weights place lower support on later mixture components than earlier ones.

By assigning a DP prior to the mixing distribution, $G(\theta)$, the DP process can be used to construct infinite mixture models. In general, the random density function for a DP mixture model is given by

$$f_G(y) = \int f(y \mid \theta) \mathrm{d}G(\theta),$$
$$G \sim DP(\alpha, G_0).$$

Given the discrete nature of the mixing distribution, which is shown through the stick-breaking construction of the mixing distribution, the random density function can also be expressed as

$$f_G(y) = \sum_{h=1}^{\infty} f(y \mid \theta_h) \omega_h, \tag{2.2}$$

where $f(y \mid \theta_h)$ is the kernel, which is a probability density function (Ghosh & Ramamoorthi, 2003), or probability mass function (PMF) for discrete data as is the case in this study. The discrete nature of the DP model leads to natural clustering of persons, which could be useful when exploring the number of clusters of persons present in the data (i.e., number of latent groups), though additional challenges arise given that there is an infinite number of mixture components to which a person could be assigned. The manner best suited to identify the distinct number of latent classes that contribute to DSF in an item is discussed in more detail below.

Dependent Dirichlet Process Rating Model

By allowing $G \sim \mathrm{DP}(\alpha, G_0)$, the parametric assumption in the form of *G* is overcome. However, this specification is still quite limiting in that the mixture distribution is assumed to be the same for all levels, or across the range, of covariates, with the covariates consisting of person, item, and/or test instrument structure information. Allowing the mixture distribution to be dependent on covariates, that is, $G_{\mathbf{x}} \sim DP(\alpha_{\mathbf{x}} G_{0\mathbf{x}})$, is generally referred to as a dependent

Dirichlet process. The model I present in this section assigns a multiple DP prior (mDP; Basu, 2007) on the random distribution *G*. The conditional random density for the model I present is given by

$$f_{G_{\mathbf{x}}}\left(y_i \mid \theta\right) = \int f\left(y_i \mid \theta, \boldsymbol{\tau}\right) \mathrm{d}G_{\mathbf{x}}\left(\boldsymbol{\tau}\right)$$

with

$$G_{\mathbf{x}} \sim \mathrm{mDP}\left(\alpha_{\mathbf{x}}, G_{0\mathbf{x}}\right)$$

The random distribution is formed via a stick-breaking process (Sethuraman, 1994):

$$G_{\mathbf{x}}\left(\cdot\right) = G_{\mathbf{x}}\left(\cdot; \boldsymbol{\tau}\left(\mathbf{x}\right), \boldsymbol{\upsilon}\left(\mathbf{x}\right)\right) = \sum\nolimits_{h=1}^{\infty} \omega_h\left(\mathbf{x}\right) \delta_{\boldsymbol{\tau}_h\left(\mathbf{x}\right)}\left(\cdot\right)$$

where $\delta\left(\cdot\right)$ is a point-mass distribution, and the covariate-dependent mixture weights are formed by

$$\omega_h\left(\mathbf{x}\right) = \upsilon_h\left(\mathbf{x}\right) \prod\nolimits_{z=1}^{h-1} \left(1 - \upsilon_z\left(\mathbf{x}\right)\right)$$

which are based on the beta random variables $\upsilon_h\left(\mathbf{x}\right) \middle| \alpha_{\mathbf{x}} \sim_{ind} \mathrm{beta}\left(1, \alpha_{\mathbf{x}}\right)$ and atoms $\boldsymbol{\tau}_h\left(\mathbf{x}\right) \middle| \boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathbf{x}} \sim_{ind} \mathrm{normal}_m\left(\boldsymbol{\mu}_{\mathbf{x}}, \Sigma_{\mathbf{x}}\right)$ for $h = 1, 2, \ldots$. The stick-breaking process reveals the discrete nature of the random distribution, which allows the DDP-RM model presented in this section to be re-expressed as a covariate-dependent infinite-mixture model. That is,

$$f_{G_{\mathbf{x}}}\left(y_i \mid \theta\right) = \sum_{h=1}^{\infty} f\left(y_i \mid \theta, \boldsymbol{\tau}_h\left(\mathbf{x}\right)\right) \omega_h\left(\mathbf{x}\right) \tag{2.3}$$

with the kernel $f\left(y_i \mid \theta, \boldsymbol{\tau}_h\left(\mathbf{x}\right)\right)$ dependent on the ability, θ, and on ordered item category steps $\boldsymbol{\tau}\left(\mathbf{x}\right) = \left(\tau_1, \ldots, \tau_m\right)'$ for all $\mathbf{x} \in \mathcal{X}$, where the response categories range from 0 to *m* at each **x**.

When the abilities are unidimensional, $\boldsymbol{\theta} = \theta_{t(i)}$ for the t^{th} person. When the abilities are Q-dimensional, $\boldsymbol{\theta} = \theta_{t(i)q(i)}$.

To approximate the covariate-dependent infinite-mixture model presented in Equation 2.3, I specified a truncated DP version of the model:

$$f_{G_{\mathbf{x}}}\left(y_i \mid \theta\right) = \sum_{h=1}^{N_{\max}=50} f\left(y_i \mid \theta, \boldsymbol{\tau}_h\left(\mathbf{x}\right)\right)\omega_h\left(\mathbf{x}\right)$$

where the number of mixture components was set to 50, $N_{\max} = 50$ for all analyses. A truncated version was selected for computational tractability. Moreover, a truncated DP has been shown to approximate the true DP (Ishwaran & James, 2001; Muliere & Tardella, 1998), given that a large enough value for $N_{\max}$ is chosen. In a DSF analysis, a value of 50 should be more than large enough because DSF often occurs across a few groups. Thus, 50 mixture components should be sufficient to form the density describing how the respondents are using the rating categories.

The following specifications for the prior distributions completed the model:

$$\begin{aligned}
&\upsilon_h\left(\mathbf{x}\right) \sim_{ind} \text{beta}\left(1, \alpha = 2\right),\ h = 1,2,\ldots, \text{ and } \mathbf{x} \in \mathcal{X};\\
&\boldsymbol{\theta} \mid \Sigma_{\boldsymbol{\theta}} \sim_{i.i.d.} \text{normal}_Q\left(\mathbf{0}, \Sigma_{\boldsymbol{\theta}}\right) \text{and } \Sigma_{\boldsymbol{\theta}} \sim \text{iw}\left(4, .1\mathbf{I}_Q\right) \text{ when } Q > 1, \text{ and}\\
&\boldsymbol{\theta} \mid \sigma^2 \sim_{i.i.d.} \text{normal}\left(0, \sigma^2\right) \text{and } \sigma^2 \sim \text{ig}\left(.01, .01\right) \text{ when } Q = 1;\\
&\boldsymbol{\tau}_h\left(\mathbf{x}\right) \mid \boldsymbol{\mu}_{\mathbf{x}} \sim_{ind} \text{normal}_m\left(\boldsymbol{\mu}_{\mathbf{x}}, .5\mathbf{I}_m\right),\ \mathbf{x} \in \mathcal{X};\\
&\boldsymbol{\mu}_{\mathbf{x}} \sim \text{normal}_m\left(0, 25\mathbf{I}_m\right),\ \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

The prior distributions assigned for all parameters were proper (i.e., distributions that integrate to 1). When **x** are item indicators (i.e., a dummy code for each item), the subscript **x** in the above expressions can be replaced with j (where $j = 1,\ldots,J$). In this study, **x** represented the items. However, to keep the expressions general in this section, I retained the **x** subscript when discussing the model in its general form.

The above prior distributions were strategically chosen. Because IRT models are naturally over-parameterized, the location of the model is not set. That is, any constant could be added to θ and τ, and the sample probability of an outcome would remains the same. Thus, a reference location is required. The location of the model can be set by fixing the mean of the prior distribution for the person abilities to zero (Patz & Junker, 1999), though this value could be any value. In order to remain consistent with the spirit of IRT modeling, I set the value to 0 for the unidimensional case and a vector of 0s for the multidimensional case. The variance of the ability distribution was freely estimated, with a vague but proper prior assigned so as to allow the data to drive the prior variance-covariance matrix estimate (for two-dimensional ability parameter) or the variance (for unidimensional ability parameter). Because fixing the mean of the prior distribution of the abilities set the location, the vector of means for the category steps could be freely estimated, and a vague but proper prior was assigned to allow the data to determine the estimates of these means. The variance-covariance matrix was selected for the category steps based on the starting assumption of low support on the category steps having DSF, but not setting the variance so small that DSF could not emerge. Finally, the α parameter was set to 2 for all analyses. The smaller this value, the greater the support for fewer numbers of mixture components corresponding to lower component indices. Given that DSF often occurs across a few subgroups (e.g., gender and race), a low value was appropriate. Other values were investigated, but larger values did not lead to greater gains.

Chapter 1 explained that cumulative probability models and adjacent-category probability models are just two of the general classes of possible IRT models for rating scale data. Each of these types of probability models, as well as others not mentioned in Chapter 1, can represent the kernel; that is, when the kernel is a PMF in the form of cumulative probabilities,

$$f\left(y_i \mid \boldsymbol{\theta}, \boldsymbol{\tau}_h\left(\mathbf{x}\right)\right) = \begin{cases} H_k\left(\eta\right), \text{ when } k = m \\ H_k\left(\eta\right) - H_{(k+1)}\left(\eta\right), \text{ when } k \neq 0 \text{ or } m \\ 1 - \sum_{k=1}^{m} P\left(Y_i = k \mid \boldsymbol{\theta}, \boldsymbol{\tau}_h\left(\mathbf{x}\right)\right), \text{ when } k = 0 \end{cases},$$

and when the kernel is a PMF representing adjacent category probabilities,

$$f\left(y_i \mid \boldsymbol{\theta}, \boldsymbol{\tau}_h\left(\mathbf{x}\right)\right) = \frac{\prod_{x=0}^{k} H_x\left(\eta\right) \prod_{z=k+1}^{m} \left[1 - H_z\left(\eta\right)\right]}{\sum_{w=0}^{m_j} \left(\prod_{x=0}^{w} H_x\left(\eta\right) \prod_{z=w+1}^{m} \left[1 - H_z\left(\eta\right)\right] \right)}.$$

In both cases, as noted in Chapter 1, $H\left(\cdot\right)$ is a CDF, and the systematic component (i.e., η) is $\eta = \theta - \tau$. For this study, the PMF representing adjacent category probabilities was assigned for the kernel, with $H\left(\cdot\right) \equiv L\left(\cdot \mid 0,1\right)$.

The parameters presented in the specification of the prior distributions represent the following:

- $\upsilon_h\left(\mathbf{x}\right)$ is the beta random variable underlying the mixture weights at **x** and is assumed to be distributed as a beta distribution with shape and scale parameters $a_h\left(\mathbf{x}\right)$ and $b_h\left(\mathbf{x}\right)$, respectively.
- $\boldsymbol{\theta}$ is the ability-level vector; that is, $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_Q\right)'$ (where q = 1, … , Q). It is assumed to be distributed as a multivariate normal with a column vector mean containing 0s and variance-covariance matrix equal to Σ_θ. When the data are assumed to measure a unidimensional trait (i.e., q = 1), the ability estimates are assumed to be normally distributed with a mean of 0 and variance equal to σ_θ^2.

- $\boldsymbol{\tau}_h(\mathbf{x})$ is the category step vector for mixture component *h* and is drawn from $G_{0\mathbf{x}}$; that is, $\boldsymbol{\tau}_h(\mathbf{x}) = (\tau_{h1},\ldots,\tau_{hm})'$ and is assumed to be distributed as a multivariate normal with a column vector mean $\mu_{\mathbf{x}}$ and variance-covariance matrix being an *m*-dimensional identity matrix with .5 on the diagonal. In this model, $\tau(\mathbf{x})$ is treated as random and is assumed to be distributed as a DP, with $G_{0\mathbf{x}} = \text{normal}_m(\mu_{\mathbf{x}}, .5\mathbf{I}_m)$.

Among the parameters described, the effects of item *j* are captured in $\boldsymbol{\tau}(\mathbf{x})$. The column vector $\boldsymbol{\tau}(\mathbf{x})$ contains the values for the relative category steps 1 to *m* (given that the response options range from 0 to *m*) at each $\mathbf{x} \in \mathcal{X}$. The mean of the posterior distribution for $\boldsymbol{\tau}(\mathbf{x})$ is analogous to item category step point estimates (within a frequentist framework) in traditional IRT models for rating-scale data in which the link functions are defined in terms of adjacent categories.

What differentiates this model from the lDP version of the DDP-RM is that, with the lDP, dependency among the mixing distribution across the covariates can exist, that is, information can be shared across covariates. The mDP version of the DDP-RM assumes the mixing distribution at each **x** is independent, so that the mixing distributions do not share information across **x**.

MCMC Algorithm

As previously noted, I relied on the Gibbs sampling approach of the truncated DP (Ishwaran & James, 2001) rather than a slice sampling approach (e.g., Kalli, Griffin, and Walker, 2011). For notational convenience, I represent the sample set of rating data by

$\mathcal{D}_n = \left\{\left(y_i, \mathbf{x}_i\right)\right\}_{i=1}^{n=NJ}$, which are provided by N persons ($t = 1,\ldots,N$) on J items, and $n = N \times J$ represents the total number of item responses in the data. Based on the Gibbs sampling approach of the truncated DP and given the cluster membership latent variable $d_i \in \left\{1,2,\ldots,N_{\max}\right\}$, the data likelihood is obtained by

$$\prod_{i=1}^{NJ} f\left(y_i \mid \boldsymbol{\theta}_{t(i)q(i)}, \boldsymbol{\tau}_{d_i}\left(\mathbf{x}_i\right)\right)\omega_{d_i}\left(\mathbf{x}_i\right). \tag{2.4}$$

Specifically, for each $i = 1,\ldots, n$ and $t = 1,\ldots, N$, each parameter value is sampled from its corresponding conditional posterior distribution at each MCMC stage:

1) $\pi\left(d_i = h \mid \ldots\right) \propto \dfrac{f\left(y_i \mid \boldsymbol{\theta}_{t(i)}, \boldsymbol{\tau}_{d_i}\left(\mathbf{x}_i\right)\right)\omega_h\left(\mathbf{x}_i\right)}{\sum_{h=1}^{\text{Nmax}} f\left(y_i \mid \boldsymbol{\theta}_{t(i)}, \boldsymbol{\tau}_{d_i}\left(\mathbf{x}_i\right)\right)\omega_h\left(\mathbf{x}_i\right)}$, $h = 1,\ldots, N_{\max}$

2) $\pi\left(\boldsymbol{\tau}_h\left(\mathbf{x}\right) \mid \ldots\right) \propto \text{normal}_{m(\mathbf{x})}\left(\boldsymbol{\mu}_{\mathbf{x}}, .5\mathbf{I}\right)\prod_{i \in h(\mathbf{x})} f\left(y_i \mid \boldsymbol{\theta}_{t(i)}, \boldsymbol{\tau}_{d_i}\left(\mathbf{x}\right)\right)$, $h = 1,\ldots, N_{\max}$

3) $\pi\left(\mu_{\mathbf{x}k} \mid \ldots\right) = \text{normal}\left(\dfrac{\left(\dfrac{\mu_0}{\sigma_0^2} + \dfrac{\sum_{i=1}^{N}\tau_{kd_i}\left(\mathbf{x}_i\right)}{\sigma_{\mathbf{x}k}^2}\right)}{\dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma_{\mathbf{x}k}^2}}, \left(\dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma_{\mathbf{x}k}^2}\right)^{-1}\right)$, where μ_0 and σ_0^2 are the mean and variance hyper priors for the distribution of τ_k, and $\sigma_{\mathbf{x}k}^2$ is the sample variance at $\mathbf{x}$

4) $\pi\left(\upsilon_h\left(\mathbf{x}\right)\middle|\ldots\right) = \text{beta}\left(a_h\left(\mathbf{x}\right) + \sum_{i=1}^{N} I\left(d_i = h\right), b_h\left(\mathbf{x}\right) + \sum_{i=1}^{N} I\left(d_i > h\right)\right)$,

where $h = 1, \ldots, N_{\max}$

When the ability distribution is unidimensional:

5a) $\pi\left(\theta_t \mid \ldots\right) \propto \text{normal}\left(0, \sigma_\theta^2\right) \prod_{i \in t} f\left(y_i \mid \theta_{t(i)}, \boldsymbol{\tau}_{d_i}\left(\mathbf{x}_i\right)\right)$

6a) $\pi\left(\sigma_\theta^2 \mid \ldots\right) = \text{ig}\left(a + \frac{N}{2}, b + \frac{\sum_{t=1}^{N} \theta_t^2}{2}\right)$, where a and b are the prior shape and rate parameters

When the ability distribution is Q-dimensional:

5b) $\pi\left(\boldsymbol{\theta}_t \mid \ldots\right) \propto \text{normal}_Q\left(\mathbf{0}, \Sigma_\theta\right) \prod_{i \in t} f\left(y_i \mid \theta_{t(i)q(i)}, \boldsymbol{\tau}_{d_i}\left(\mathbf{x}_i\right)\right)$

6b) $\pi\left(\Sigma_\theta \mid \ldots\right) = \text{iw}\left(N + v, \left(\Lambda + \sum_{t=1}^{N} \boldsymbol{\theta}_t \boldsymbol{\theta}_t'\right)^{-1}\right)$, where Λ is a prior scale matrix

Standard Gibbs sampling methods can be used for steps 1, 3, 4, and 6, while the Metropolis-Hastings algorithm can be used for the remaining steps. If all 6 steps are repeated a sufficiently large number of times (e.g., 100,000 times), samples of model parameters that converge to the posterior distribution of the high-dimensional model are obtained. The precision parameter α was fixed to 2 for this study, though it could easily be modeled by assigning a gamma distribution on α. Lower values of α lead to fewer clusters (Ishwaran & James, 2000). When examining items for DIF/DSF, one does not need a large number of clusters because DIF/DSF usually occurs across a few groups. Thus, a low value for α was reasonable.

Bayesian Posterior Inference of the DDP-RM

I denote the parameters of the mDP version of the DDP-RM with $\boldsymbol{\xi} = \left(\boldsymbol{\theta}, \Sigma_\theta, \boldsymbol{\tau}, \boldsymbol{\upsilon}, \boldsymbol{\mu}\right)$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_t)_{t=1}^{N}$, $\boldsymbol{\tau}\left(\mathbf{x}\right) = \left(\boldsymbol{\tau}_h\left(\mathbf{x}\right)\right)_{h=1, \mathbf{x} \in \mathcal{X}}^{N_{\max}}$, and $\boldsymbol{\upsilon}\left(\mathbf{x}\right) = \left(\upsilon_h\left(\mathbf{x}\right)\right)_{h=1, \mathbf{x} \in \mathcal{X}}^{N_{\max}}$. According to Bayes theorem, the posterior density of $\boldsymbol{\xi}$ is given by

$$\pi\left(\boldsymbol{\xi}\middle|\mathcal{D}_n\right)=\frac{\prod_{i=1}^{n} f_{G_{\mathbf{x}}}\left(y_i\,|\,\boldsymbol{\xi}\right)\pi\left(\boldsymbol{\xi}\right)}{\int_{\Omega_{\boldsymbol{\xi}}}\prod_{i=1}^{n} f_{G_{\mathbf{x}}}\left(y_i\,|\,\boldsymbol{\xi}\right)\mathrm{d}\Pi\left(\boldsymbol{\xi}\right)}$$

for data $\mathcal{D}_n$ with likelihood $\prod_{i=1}^{n} f_{G_{\mathbf{x}}}\left(y_i\,|\,\boldsymbol{\xi}\right)$ under a model with $\boldsymbol{\xi}$, and proper prior densities $\pi(\boldsymbol{\xi})$ defined over the space $\Omega_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$. The posterior predictive density of *Y* for a chosen **x** is given by

$$f_n\left(y\middle|\mathbf{x}\right)=\int f\left(y\middle|\mathbf{x};\boldsymbol{\xi}\right)\pi\left(\boldsymbol{\xi}\middle|\mathcal{D}_n\right)\mathrm{d}\boldsymbol{\xi}\,.$$

This density can be used to obtain the posterior predictive mean (expectation)

$\mathrm{E}_n(Y\,|\,\mathbf{x})=\int y f_n\left(y\,|\,\mathbf{x}\right)\mathrm{d}y$ and variance $\mathrm{Var}_n(Y\,|\,\mathbf{x})=\int\left\{y-E(y\,|\,\mathbf{x})\right\}^2 f_n\left(y\,|\,\mathbf{x}\right)\mathrm{d}y$.

When examining items for DSF, of primary interest is to infer functionals of the posterior predictive mean $\mathrm{E}_n\left[G_{\mathbf{x}}\left(\cdot\right)\right]$ of the mixture step distribution $G_{\mathbf{x}}\left(\boldsymbol{\tau}\right)$, such as its density. The posterior predictive mean of the mixture step distribution is defined by

$$\mathrm{E}_n\left[G_{\mathbf{x}}\left(\cdot\right)\right]=\iint G_{\mathbf{x}}\left(\cdot\,;\boldsymbol{\tau}\left(\mathbf{x}\right),\boldsymbol{\upsilon}\left(\mathbf{x}\right)\right)\pi\left(\boldsymbol{\tau}\left(\mathbf{x}\right),\boldsymbol{\upsilon}\left(\mathbf{x}\right)\middle|\,\mathcal{D}_n\right)\mathrm{d}\boldsymbol{\tau}\left(\mathbf{x}\right)\mathrm{d}\boldsymbol{\upsilon}\left(\mathbf{x}\right),$$

given the marginal posterior density

$$\pi\left(\boldsymbol{\tau}\left(\mathbf{x}\right),\boldsymbol{\upsilon}\left(\mathbf{x}\right)\middle|\,\mathcal{D}_n\right)=\int\int\int\pi\left(\boldsymbol{\theta},\Sigma_{\theta},\boldsymbol{\tau}\left(\mathbf{x}\right),\boldsymbol{\upsilon}\left(\mathbf{x}\right),\mu_{\mathbf{x}}\,|\,\mathcal{D}_n\right)\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\Sigma_{\theta}\mathrm{d}\mu_{\mathbf{x}}.$$

The shape of $\mathrm{E}_n\left[G_{\mathbf{x}}\left(\cdot\right)\right]$ indicates whether an item has DSF. Standard MCMC sampling techniques, which were previously described, were used to perform inference of functionals of $\pi\left(\boldsymbol{\xi}\,|\,\mathcal{D}_n\right)$, marginal posterior densities, posterior predictive densities $f_n\left(y\middle|\mathbf{x}\right)$, and the posterior mean mixing distribution $\mathrm{E}_n\left[G_{\mathbf{x}}\left(\cdot\right)\right]$.

Identifiability of Mixture Models

As presented in Equation 2.4, the model likelihood is based on $N_{\max}$ number of mixtures. It is well known that mixture models are unidentifiable because of the so-called label-switching problem (Diebolt & Robert, 1994; Richardson & Green, 1997), in which the likelihood function remains invariant under permutation of the density labels.

The label-switching problem, however, can be overcome in Bayesian mixture models by marginalizing out all model parameters with respect to the posterior distribution; that is, by focusing the inferences on the posterior predictive distribution of the model (Gelfand & Sahu, 1999; Poirier, 1998). Notice that the posterior predictive density follows

$$f_n\left(\boldsymbol{y}\middle|\mathbf{x}\right)=\int f\left(\boldsymbol{y}\,|\,\boldsymbol{\tau}\left(\mathbf{x}\right)\right)\pi\left(\boldsymbol{\tau}\left(\mathbf{x}\right)\middle|\mathcal{D}_n\right)\mathrm{d}\boldsymbol{\tau}\left(\mathbf{x}\right).$$

If one is indeed interested in learning about a nonidentified parameter in the posterior (e.g., the category step parameter τ), after MCMC sampling, the parameter of interest should be marginally informative. Specifically, the nonidentified parameter of interest is marginally informative when the marginal prior for the parameter is not equal to the posterior of the parameter; that is,

$$\pi\left(\boldsymbol{\tau}\left(\mathbf{x}\right)\right)\neq\pi\left(\boldsymbol{\tau}\left(\mathbf{x}\right)\middle|\mathcal{D}_n\right), \tag{2.5}$$

so that the data enable "Bayesian learning" of the model parameter. Equation 2.5 will be true when the prior distributions for the parameters are proper (Gelfand & Sahu, 1999; Poirier, 1998). Moreover, as presented in the section on MCMC sampling methods, for each model parameter, the full conditional posterior distribution depends on the data. Therefore, when posterior inference is undertaken for a model parameter of interest, by taking marginal MCMC posterior samples of the parameter, the marginal prior for the parameter is data-dependent and therefore different from the marginal prior of the parameter.

Resulting Information

Analysis of the data with the DDP-RM can provide information analogous to that provided with more common IRT models, such as ability estimates for each person and category step estimates, although these estimates are means of the posterior distributions. To obtain the 95% Monte Carlo (MC) confidence intervals for each estimate of the posterior mean parameter, I relied on the batch mean analysis approach described by Flegal and Jones (2011).

In addition to information similar to that of traditional IRT models, the DDP-RM provides the number of latent classes present in the data for a given item. The number of latent classes is ascertained through the number of modes in the posterior predictive density of the category thresholds for the item of interest, if no other information is contained in the **x** in $G_{\mathbf{x}}$ other than item indicators.

Unique Features of the Model

One approach to examining whether an item category step has DSF is to examine the posterior predictive mean density of the mixture step distribution, $\mathrm{E}_n\left[G_{\mathbf{x}}(\cdot)\right]$. When a step is free of DSF, the density corresponding to that step should be unimodal with very small variance, which indicates that only one unique step value is needed to describe the data corresponding to that step. When a step has DSF between two groups, the density should be bimodal, with each mode corresponding to a value for a group. By subjecting the category steps to an infinite number of mixtures, which the DDP-RM does, all forms for the step densities are supported, including various bimodal densities.

With a parametric IRT model, when the category steps are treated as random across persons, the true form for the step distribution is assumed to follow a normal distribution, with

mean and variance estimated from the data. This prior would be appropriate for category steps that are free of DSF, but inappropriate when a step has DSF between two groups, which would require a bimodal form for the distribution.

The covariate-dependent mixing distribution, $G_{\mathbf{x}}$, based on an infinite number of mixtures, allows the density to take on any form. The form could change across items when **x** are item indicators, which is the case for this study. Thus, the mixing distribution of the mDP model is applicable when some items have steps with DSF (requiring multimodal form for the mixing distribution) while other items' steps are free of DSF (requiring unimodal form for the mixing distribution).

Data Sets Analyzed

Simulated Data

To evaluate the DDP-RM's ability to recover the number of generating latent classes (i.e., latent clustering configuration) contributing to DSF in a subset of items, I conducted a simulation study in which the presented model was used to analyze generated data. Three-point rating scale data ($k = 0, 1, 2$) were generated for 20 items. The simulation study consisted of a 2 × 2 × 5 non-nested design, which was as follows: *number of persons* (400 or 800) by *number of dimensions* (unidimensional or two dimensional) by *clustering configuration*. The levels for *clustering configuration* varied depending on the dimensionality of the data. Within each condition, a single data set was generated based on a set of generating values for that condition. For the conditions corresponding to $N = 400$, $n = N \times J = 400 \times 20 = 8{,}000$ total rating observations were generated, and for the conditions corresponding to $N = 800$, 16,000 total rating observations were generated.

For the unidimensional and two-dimensional conditions, when the clustering condition for persons called for two groups, each group had 200 or 400 persons, depending on whether the total number of persons was 400 or 800, respectively. When the data generating condition involved unidimensional examinee abilities (Conditions 1 through 5), the abilities were drawn from a normal distribution with a mean of 0 and a variance of 1. When the data generating condition involved two-dimensional person abilities (Conditions 6 through 10), the abilities were drawn from a multivariate normal with a vector of 0s and a variance-covariance matrix with variances of 1 and a correlation of .5. Moreover, within each ability dimensional condition, the first three levels represented non-DSF conditions (Conditions 1 through 3 and 6 through 8), and the last two levels represented DSF conditions (Conditions 4, 5, 9, and 10).

Unidimensional ability condition. The five clustering configuration levels (i.e., conditions 1-5) for the unidimensional condition were as follows:

1) A single cluster of items and group of persons, with a single category step structure applying to all items. This condition represents a rating scale model (RSM).
2) Twenty clusters of items, with each item having its own category step structure, and only one cluster of persons. The relative category steps were randomly drawn from a uniform distribution with range (–.3, .3). This condition represents a PCM.
3) Three clusters of items (the first and second clusters consisting of the first seven and next seven items, respectively, and the last cluster consisting of the remaining six items) and a single cluster of persons. This condition falls between the RSM and the PCM and is referred to as a blocked RSM.
4) Two clusters of persons, with three clusters of items forming in each group. The first two clusters of items (7 items in each cluster) were the same between the two groups, but the

third cluster (6 items) differed and represented DSF in the second threshold of these items. For Group 1, the items within the third cluster were more difficult. That is, the first group's generating values for these items was 1 logit larger for the second step than the second group's values. This clustering configuration represents a DSF scenario because the two groups did not share the same generating second category step for the third cluster of items.

5) Two clusters of persons. As for the item-cluster configuration, the first two clusters of items (i.e., the first and second clusters consisting of the first seven and next seven items, respectively) were common between the two groups. The third cluster of items (consisting of the remaining six items) was not common between the two groups. That is, for Group 1, the generating values were ordered (i.e., the step parameter values monotonically increased with the rating categories) for these item steps but were disordered for Group 2. Disordering occurs when the step estimate for the higher category is less than the step estimate for the lower category. The third cluster of items, then, represents a DSF condition, because the set of items did not share the same item category step parameters between the two groups.

Two-dimensional ability condition. In a two-dimensional condition, the first 10 items were indicators for the first dimension, and the remaining 10 items were indicators for the second dimension. The following were the 5 levels of cluster configurations for the two-dimensional condition (Conditions 6-10):

6) A single cluster of persons and items in each of the two dimensions. This condition represents a two-dimensional RSM.

7) A single cluster of persons and 10 clusters of items in each of the dimensions. The relative category steps were randomly drawn from a uniform distribution with range (–.3, .3). This condition represents a two-dimensional PCM.
8) A single cluster of persons and two clusters of items in each of the two dimensions. For the item configuration, the first five and second five items in each dimension formed the first and second clusters, respectively. This condition represents a two-dimensional, four-blocked RSM.
9) Two clusters of persons and items in each dimension. In both dimensions, each cluster of items consisted of five items. For the first dimension, the first cluster of items was free of DSF and DIF; the second cluster of items had DSF in the second category step with 1-logit difference in magnitude. For the second dimension, the first cluster of items was free of DSF and DIF; the second cluster of items had DSF in the first category step with 1-logit difference in magnitude. For all DSF items, the generating step value for the DSF step was greater for the first group than the value for the second group.
10) The first dimension consisted of one cluster of persons and items for both groups. Thus, the items in dimension 1 were free of DSF. In the second dimension, there were three clusters of items for each of the two clusters of persons, with the item-cluster configuration varying across the groups. Within this dimension, four items had .5-logit DSF effect in the first step and 1-logit DSF effect in the second step, and two items had .5-logit DIF between the two groups (i.e., the first and second steps were .5-logit greater for the first group).

The cluster configuration for the unidimensional and two-dimensional conditions was repeated for each of the two sample size conditions. Table II provides a brief overview of the

cluster configurations by dimensions. The generating parameter values for Conditions 1 through 3 are in Table III. The generating parameter values for Conditions 4 and 5 are in Table IV and V, respectively. The generating parameter values for Conditions 6 through 8 are in Table VI. The generating parameter values for Conditions 9 and 10 are in Table VII and VIII, respectively. The values for all but Conditions 2 and 7 were based on values selected to target the ability distribution(s). Conditions 2 and 7 had targeted overall item difficulties that were used in all other conditions, but the relative category steps were randomly generated from a uniform distribution with range (–.3, .3). For these two conditions, the randomly drawn values for the first and second category steps were subtracted and added, respectively, from the overall item difficulty value.

For all conditions, the baseline IRT model was one of the Rasch models for rating-scale data (i.e., RSM, PCM, or blocked RSM), because the purpose of this study was to present and explore a model in which DSF could be identified through the marginal posterior mean density estimates for the category step mixing distributions. To prevent varying item discrimination levels from confounding the results of item clustering identification based on category steps, I set the data-generating model to one in which all items shared a common discrimination level.

For conditions in which all items belonged to a single cluster and all persons belonged to a single group, the data were generated under the RSM. For conditions in which subsets of items belonged to different clusters, the data were generated under a blocked RSM (i.e., a model in which subsets of items share similar category step parameter structure but more than a single cluster of items exists). For conditions that had two groups of persons, the data were generated in a manner in which the item characteristics for each group conformed to one of the Rasch models

TABLE II

CLUSTERING CONFIGURATION FOR EACH OF THE DIMENSIONAL CONDITIONS

		Number of Dimensions	
		Unidimensional	Two-Dimensional
Cluster Configuration	No DSF	Single cluster of items and group of persons (i.e., RSM)	Single cluster of items for each of the two dimensions; single cluster of persons (i.e., two-dimensional RSM)
		Twenty clusters of items; single cluster of persons (i.e. PCM)	Twenty clusters of items; single cluster of persons (i.e. two-dimensional PCM)
		Three clusters of items; single cluster of persons	Two clusters of items in the first dimension; two clusters of items in the second dimension; single cluster of persons
	DSF	Three clusters of items for each of the two clusters of persons. The third cluster had DSF.	Two clusters of items and persons for each of the two dimensions. Within each dimension, one cluster of items had DSF.
		Three clusters of items for each of the two clusters of persons. The third cluster of items had DSF; that is, the generating values were disordered for Group 2.	Dimension 1: One cluster of items and persons (no DSF). Dimension 2: Three clusters of items for each of the two clusters of persons. Four items had DSF, and two items had DIF.

Note. The cluster configurations by dimensionality were repeated for each of the sample size conditions ($N = 400$ or $N = 800$).

TABLE III

GENERATING CATEGORY THRESHOLD VALUES FOR CONDITIONS 1 THROUGH 3

	Condition					
	1		2		3	
Item	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
1	–2.00	–1.00	–1.81	–0.91	–1.75	–1.25
2	–1.84	–0.84	–1.60	–1.12	–1.84	–0.84
3	–1.68	–0.68	–1.91	–0.47	–1.93	–0.43
4	–1.53	–0.53	–1.28	–0.27	–1.28	–0.78
5	–1.37	–0.37	–1.29	–0.26	–1.37	–0.37
6	–1.21	–0.21	–1.45	–0.06	–1.46	0.04
7	–1.05	–0.05	–1.19	0.09	–0.80	–0.30
8	–0.89	0.11	–0.87	0.04	–0.89	0.11
9	–0.74	0.26	–0.46	0.36	–0.99	0.51
10	–0.58	0.42	–0.30	0.22	–0.33	0.17
11	–0.42	0.58	–0.63	0.70	–0.42	0.58
12	–0.26	0.74	0.02	0.46	–0.51	0.99
13	–0.11	0.89	0.17	0.76	0.14	0.64
14	0.05	1.05	0.04	0.78	0.05	1.05
15	0.21	1.21	0.39	0.97	–0.04	1.46
16	0.37	1.37	0.15	1.56	0.62	1.12
17	0.53	1.53	0.48	1.64	0.53	1.53
18	0.68	1.68	0.93	1.57	0.43	1.93
19	0.84	1.84	1.02	2.11	1.09	1.59
20	1.00	2.00	1.28	1.72	1.00	2.00

Note. In these conditions, all items were free of DSF. Thus, only one set of parameters was required for all persons.

TABLE IV

GENERATING CATEGORY THRESHOLD VALUES FOR CONDITION 4 FOR EACH OF THE TWO GROUPS AND THE DSF EFFECT

	Group					
	1		2		DSF Effect	
Item	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
1	–2.00	–1.00	–2.00	–1.00		
2	–1.84	–1.09	–1.84	–1.09		
3	–1.68	–0.43	–1.68	–1.43	0.0	1.0
4	–1.53	–0.53	–1.53	–0.53		
5	–1.37	–0.62	–1.37	–0.62		
6	–1.21	0.04	–1.21	–0.96	0.0	1.0
7	–1.05	–0.05	–1.05	–0.05		
8	–0.89	–0.14	–0.89	–0.14		
9	–0.74	0.51	–0.74	–0.49	0.0	1.0
10	–0.58	0.42	–0.58	0.42		
11	–0.42	0.33	–0.42	0.33		
12	–0.26	0.99	–0.26	–0.01	0.0	1.0
13	–0.11	0.89	–0.11	0.89		
14	0.05	0.80	0.05	0.80		
15	0.21	1.46	0.21	0.46	0.0	1.0
16	0.37	1.37	0.37	1.37		
17	0.53	1.28	0.53	1.28		
18	0.68	1.93	0.68	0.93	0.0	1.0
19	0.84	1.84	0.84	1.84		
20	1.00	1.75	1.00	1.75		

Note. Items that do not contain values under the DSF Effect columns were free of DSF.

TABLE V

GENERATING CATEGORY THRESHOLD VALUES FOR CONDITION 5 FOR EACH OF THE TWO GROUPS AND THE DSF EFFECT

	Group					
	1		2		DSF Effect	
Item	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
1	–1.75	–1.25	–1.75	–1.25		
2	–1.84	–0.84	–1.84	–0.84		
3	–1.68	–0.68	–0.68	–1.68	–1.00	1.00
4	–1.28	–0.78	–1.28	–0.78		
5	–1.37	–0.37	–1.37	–0.37		
6	–1.21	–0.21	–0.21	–1.21	–1.00	1.00
7	–0.80	–0.30	–0.80	–0.30		
8	–0.89	0.11	–0.89	0.11		
9	–0.74	0.26	0.26	–0.74	–1.00	1.00
10	–0.33	0.17	–0.33	0.17		
11	–0.42	0.58	–0.42	0.58		
12	–0.26	0.74	0.74	–0.26	–1.00	1.00
13	0.14	0.64	0.14	0.64		
14	0.05	1.05	0.05	1.05		
15	0.21	1.21	1.21	0.21	–1.00	1.00
16	0.62	1.12	0.62	1.12		
17	0.53	1.53	0.53	1.53		
18	0.68	1.68	1.68	0.68	–1.00	1.00
19	1.09	1.59	1.09	1.59		
20	1.00	2.00	1.00	2.00		

Note. Items that do not contain values under the DSF Effect columns were free of DSF.

TABLE VI

GENERATING CATEGORY THRESHOLD VALUES FOR CONDITIONS 6 THROUGH 8

	Condition					
	6		7		8	
Item	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
1	–1.90	–0.90	–1.94	–0.75	–1.65	–1.15
2	–1.57	–0.57	–1.64	–0.71	–1.57	–0.57
3	–1.23	–0.23	–1.07	–0.23	–0.98	–0.48
4	–0.90	0.10	–0.72	0.22	–0.90	0.10
5	–0.57	0.43	–0.75	0.67	–0.32	0.18
6	–0.23	0.77	–0.24	1.04	–0.23	0.77
7	0.10	1.10	0.07	1.13	0.35	0.85
8	0.43	1.43	0.52	1.22	0.43	1.43
9	0.77	1.77	0.89	1.56	1.02	1.52
10	1.10	2.10	1.25	1.95	1.10	2.10
11	–2.20	–1.20	–2.33	–1.00	–1.95	–1.45
12	–1.87	–0.87	–1.76	–1.01	–1.87	–0.87
13	–1.53	–0.53	–1.44	–0.34	–1.28	–0.78
14	–1.20	–0.20	–1.40	–0.35	–1.20	–0.20
15	–0.87	0.13	–1.10	0.39	–0.62	–0.12
16	–0.53	0.47	–0.53	0.38	–0.53	0.47
17	–0.20	0.80	0.08	0.62	0.05	0.55
18	0.13	1.13	0.04	0.98	0.13	1.13
19	0.47	1.47	0.52	1.54	0.72	1.22
20	0.80	1.80	0.63	1.78	0.80	1.80

TABLE VII

GENERATING CATEGORY THRESHOLD VALUES FOR CONDITION 9 FOR EACH OF THE TWO GROUPS AND THE DSF EFFECT

	Group					
	1		2		DSF effect	
Item	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
1	–1.90	–0.90	–1.90	–0.90		
2	–1.57	–0.07	–1.57	–1.07		1.00
3	–1.23	–0.23	–1.23	–0.23		
4	–0.90	0.60	–0.90	–0.40		1.00
5	–0.57	0.43	–0.57	0.43		
6	–0.23	1.27	–0.23	0.27		1.00
7	0.10	1.10	0.10	1.10		
8	0.43	1.93	0.43	0.93		1.00
9	0.77	1.77	0.77	1.77		
10	1.10	2.60	1.10	1.60		1.00
11	–2.20	–1.20	–2.20	–1.20		
12	–1.37	–0.87	–2.37	–0.87	1.00	
13	–1.53	–0.53	–1.53	–0.53		
14	–0.70	–0.20	–1.70	–0.20	1.00	
15	–0.87	0.13	–0.87	0.13		
16	–0.03	0.47	–1.03	0.47	1.00	
17	–0.20	0.80	–0.20	0.80		
18	0.63	1.13	–0.37	1.13	1.00	
19	0.47	1.47	0.47	1.47		
20	1.30	1.80	0.30	1.80	1.00	

TABLE VIII

GENERATING CATEGORY THRESHOLD VALUES FOR CONDITION 10 FOR EACH OF THE TWO GROUPS AND THE DSF EFFECT

	Group					
	1		2		DSF effect	
Item	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
1	–1.90	–0.90	–1.90	–0.90		
2	–1.57	–0.57	–1.57	–0.57		
3	–1.23	–0.23	–1.23	–0.23		
4	–0.90	0.10	–0.90	0.10		
5	–0.57	0.43	–0.57	0.43		
6	–0.23	0.77	–0.23	0.77		
7	0.10	1.10	0.10	1.10		
8	0.43	1.43	0.43	1.43		
9	0.77	1.77	0.77	1.77		
10	1.10	2.10	1.10	2.10		
11	–2.20	–0.95	–2.70	–1.95	0.50	1.00
12	–1.87	–1.12	–1.87	–1.12		
13	–1.53	–0.28	–2.03	–1.28	0.50	1.00
14	–1.20	–0.45	–1.70	–0.95	0.50	0.50
15	–0.87	–0.12	–1.37	–0.62	0.50	0.50
16	–0.53	0.22	–0.53	0.22		
17	–0.20	1.05	–0.70	0.05	0.50	1.00
18	0.13	0.88	0.13	0.88		
19	0.47	1.72	–0.03	0.72	0.50	1.00
20	0.80	1.55	0.80	1.55		

Note. Items that do not contain values under the DSF Effect columns were free of DSF.

(referred to as a multiple-group RSM or blocked RSM). When DSF items were present, the DSF effect was .5- or 1-logit difference between the two groups, which represented either half or a full population standard deviation. The mean of the distribution from which the two groups were drawn was the same (i.e., the mean for each group was 0). For two-dimensional conditions, the data were generated under a two-dimensional Rasch model, and the ability estimates between the two dimensions had a correlation of .50.

Real-life Data

I applied the DDP-RM to two real-life data sets: the verbal aggression data (De Boeck & Wilson, 2004) and the acculturative family distancing (AFD) data (Hwang, Wood, & Fujimoto, 2010).

Verbal aggression data set. This data set contains item responses from 316 students (243 females and 73 males) from a Dutch-speaking Belgian university. The students rated 24 items, which are indicators for levels of verbal aggression (e.g., "A bus fails to stop for me. I would want to curse."), on a scale of 0 = no, 1 = perhaps, and 2 = yes. The items can be categorized into a 2 × 2 × 3 structure: *Behavior Mode* (Want or Do) by *Situation Type* (Other-to-blame or Self-to-blame) by *Behavior Type* (Curse, Scold, or Shout). Table IX provides a breakdown on the number of items within each category. The *Behavior Mode* was used to distinguish two-dimensionality, similar to the approach used by De Boeck and Wilson (2004). Thus, all models applied to this data set were two-dimensional.

AFD data set. The AFD questionnaire is composed of 46 items and measures two dimensions (i.e., *Communication Barrier* and *Values Incongruence*) of AFD. For this study, I analyzed the data corresponding to the 24 items that comprise the *Communication Barrier* dimension. Thus, all models applied to this data set were unidimensional.

TABLE IX

NUMBER OF ITEMS BY BEHAVIOR MODE, SITUATION TYPE, AND BEHAVIOR TYPE

Behavior Mode	Situation Type	Behavior			Column Total
		Curse	Scold	Shout	
Want	Other-to-blame	2	2	2	6
	Self-to-blame	2	2	2	6
Do	Other-to-blame	2	2	2	6
	Self-to-blame	2	2	2	6
Row Total		8	8	8	24

The data corresponding to 293 participants (102 mothers, 79 fathers, and 112 children of the mothers and/or fathers) who provided responses to all items were analyzed for this study. All participants were of Asian descent. Originally, on a rating scale ranging from 1 = strongly disagree to 7 = strongly agree, the participants rated their level of agreement with various statements about the level of communication between parent and child (e.g., "If my parent[s] communicated emotional distress through physical symptoms, I would understand what the physical symptoms meant" when the children completed the questionnaire, and the term "parent[s]" was replaced with "child" when the parents completed the questionnaire). A previous study suggested that the rating categories should be collapsed (Hwang, Wood, & Fujimoto, 2010). Thus, for this study, the three levels of disagreement categories were rescored to a value of 0, the neutral category was rescored to 1, and the three levels of agreement categories were rescored to 2.

All children attended the same high school, which was located in the western United States. The subsample of children consisted of 55% females and 45% males and varied from freshmen to seniors in terms of grade levels. The mothers were born in a variety of countries, such as Taiwan, mainland China, Hong Kong, Vietnam, Burma, and Thailand. The fathers were also born in a variety of countries, such as Taiwan, mainland China, Hong Kong, Vietnam, and Burma.

Evaluating the New Model

In this study, I compared the performance of the new model to traditional IRT models (i.e., RSM, PCM, generalized partial credit [GPCM], graded response model [GRM], modified graded response model [MGRM], finite-mixture Rasch model, and nominal response model [NRM]) with respect to ability parameter recovery, item category parameter recovery, and predictive performance. The comparison in model performance in terms of ability and item category step parameter recovery were assessed for only the generated data sets. The predictive performance of all models was assessed based on the analysis of the generated and real-life data sets. I assessed the number of latent class recovery only for the DDP-RM, on only the generated data sets.

The mDP model introduced in this section was coded and fit in C++, and all traditional IRT models except the mixed Rasch model were fit to the data using flexMIRT (Cai, 2012). The mixed Rasch model was fit to the data using WINMIRA (von Davier, 2001) and only for the unidimensional condition because of software limitations.

Recovery of Person Abilities and Item Category Steps

Recovery of person ability. To determine how well the new model and the various comparison IRT models recovered the generating person abilities, I relied on the root mean squared deviation (RMSD), also referred to as root mean squared error:

$$RMSD_{\theta} = \sqrt{\frac{\sum_{q=1}^{Q}\sum_{t=1}^{N}\left(\theta_{tq} - \theta'_{tq}\right)^2}{N \times Q}},$$

where θ_{tq} is the t^{th} person's estimated ability level in dimension q, and θ'_{tq} is the generating ability level value for person t in dimension q (where $t = 1, 2, \ldots, N$, with N representing the total sample size; and $q = 1,\ldots, Q$, with Q representing the number of dimensions). Values closer to 0 suggest better recovery of the generating person ability values.

For each of the generated data sets within each condition, an $RMSD_{\theta}$ was calculated using the ability estimates yielded from the DDP-RM and the comparison IRT models. Within each condition, I compared the $RMSD_{\theta}$ associated with each model. The model associated with the lowest $RMSD_{\theta}$ value was deemed the best-performing model in terms of ability-parameter recovery. The mDP model was expected to have lower $RMSD_{\theta}$ than the comparison models in conditions in which Items have DSF (Conditions 4, 5, 9, and 10) because the model was expected to control for the DSF when estimating the ability parameters.

Recovery of item category steps. To determine how well the DDP-RM and the various comparison IRT models recovered the generating item category step values, I calculated the RMSD between the estimated and the generating category step values for all items within each simulation condition:

$$RMSD_{\tau} = \sqrt{\frac{\sum_{j=1}^{J}\sum_{k=1}^{m}\sum_{t=1}^{N}\left(\tau_{jkt} - \tau'_{jkt}\right)^2}{J \times m \times N}},$$

where τ_{jkt} is the estimated k^{th} category step (where $k = 0, 1, 2$) and τ'_{jkt} is the generating value for the k^{th} relative category step for item j and person t. For most of the comparison IRT models (with the exception of the finite-mixture Rasch model), all persons had the same value for the k^{th} relative step for item j (i.e., $\tau_{jkt} = \tau_{jk}$ for all $t = 1, \ldots, N$) because these IRT models treat the step parameters as fixed effects. As such, all persons shared the same estimated value for each k^{th} step for item j. The finite-mixture Rasch model can estimate different τ_{jk} for different latent classes of persons. Thus, the persons within each latent class had the same value for each of the k^{th} relative step for item j, but persons belonging to separate clusters did not have the same value. The DDP-RM estimated a value for τ_{jk} for each person. That is, it was possible that $\tau_{jkt} \neq \tau_{jkt'}$ for all persons $t = 1,\ldots, N$ and $t \neq t'$.

By summing across all persons the difference between the estimated and generating value for the k^{th} step for item j, an RMSDτ was calculated, even though all models did not estimate the same number of values for each τ_{jk}. Lower values suggest better recovery of the generating relative category step values. The mDP model was expected to have lower RMSDτ values than the comparison models in conditions in which items have DSF (Conditions 4, 5, 9, and 10) because the model was expected to detect the DSF, thereby providing more accurate estimates of the category steps.

For each generated data set within each condition, an RMSDτ was calculated for each model, and the values across the models were compared. The model associated with the lowest

RMSDτ was considered to be the best at recovering the generating item category step parameters values.

The RMSD involving the category steps was only calculated for models in which the link functions were defined in terms of adjacent categories, which included the mDP, finite-mixture PCM, GPCM, PCM, and the RSM**.** The step parameters within the NRM, which relies on baseline category probabilities, and the GRM, which defines the link functions in terms of cumulative probabilities, have different interpretation from step parameters when the link functions are defined in terms of adjacent category probabilities. Thus, comparing the step estimates from models with different types of link functions would not be beneficial.

Predictive Performance of the Data

I compared the predictive performance of the generated and real-life data of the mDP model against the comparison IRT models, using the mean-squared predictive criterion $D(m)$ covered in Chapter I. As previously noted, the $D(m)$ is given by the following when the parameters of model $\underline{m}$ are estimated under Bayesian inference:

$$\begin{aligned} D\left(\underline{m}\right) &= \sum_{i=1}^{n}\left[y_i - \mathrm{E}_n\left(Y_i \mid \underline{\mathbf{x}}_i, \underline{m}\right)\right]^2 + \sum_{i=1}^{n} \mathrm{Var}_n\left(Y_i \mid \mathbf{x}_i \underline{m}\right) \\ &= \mathrm{GF}\left(\underline{m}\right) + \mathrm{Pen}\left(\underline{m}\right). \end{aligned}$$

The $D(\underline{m})$ can be calculated for a non-Bayesian model having point estimate $\hat{\varphi}_n = \hat{\varphi}\left(\mathcal{D}_n\right)$, such as a maximum-likelihood estimate. The $D(\underline{m})$ criterion, in this case, is estimated via $\hat{\mathrm{E}}\left(Y_z \mid \mathbf{x}_z, \underline{m}\right) = \mathrm{E}\left(Y_z \mid \mathbf{x}_z, \underline{m}, \hat{\varphi}\right)$ and $\widehat{\mathrm{Var}}\left(Y_z \mid \mathbf{x}_z, \underline{m}\right) = \mathrm{Var}\left(Y_z \mid \mathbf{x}_z, \underline{m}, \hat{\varphi}\right)$ (for $z = 1, 2, \ldots, n$, where n is the total number of observations in data set $\mathcal{D}_n$) (Gelfand & Ghosh, 1998).

Because the $D(m)$ is based on the expected value of y and variance of Y given model $\underline{m}$, this criterion can be used to compare models in which the parameters are estimated under a Bayesian inference and frequentist approach. This fact is critical because the comparison models estimate the parameters using a frequentist approach (i.e., MMLE for all comparison models except the mixed Rasch model, which uses CMLE), whereas the DDP-RM estimates the parameters under Bayesian inference. In this case, a predictive performance index that can compare the predictive fit to the data across all models in which the parameters were estimated under different approaches was required for this study, which the $D(\underline{m})$ fulfills.

For each generated data set within each condition included in the simulation study and each real-life data set, the $D(\underline{m})$ was calculated based on the set of parameter estimates from each comparison model and the DDP-RM. Within each condition in the simulation study and for each real-life data set, the $D(\underline{m})$s were compared to assess all models' performance in terms of their predictive performance, with lower values suggesting better performance. The index associated with the model presented in this chapter was expected to have the lowest value in each of the conditions in the simulation study and for each of the real-life data sets.

Proceeding Chapters

The motivation for developing the mDP model was to address the shortcomings of other IRT models currently employed in DSF analysis, which were discussed in Chapter I. Most notable of these limitations are that the true distribution of the random effects, G, follows a parametric form; the same latent class structures are assumed to apply across all items and dimensions (if multiple dimensions are specified); and some type of decision on establishing a common metric across the groups is required to examine the items for DSF, with finding a subset of items to serve as anchors as the optimal method.

The DDP-RM overcomes these limitations. With this model, an mDP prior is assigned to *G*, which supports all discrete distributions. Moreover, this prior allows for the mixture distributions for the item category steps to vary across covariates. The mDP model also allows for the number of latent groups to vary across items and dimensions of a latent construct measured by the test items. Finally, the need to find items to serve as anchors is circumvented because all persons are assumed to come from a common distribution.

In Chapter III, I present the results of the simulation study. In Chapter IV, I describe applications of the proposed model to real rating data. In both Chapters III and IV, I compare the predictive performance of the proposed model against standard IRT models. In Chapter V, I provide some additional follow-up analyses to explain the findings in Chapters III and IV. In Chapter VI, I conclude with a discussion of practical implications of the DDP-RM, as well as the model's limitations and possible future modeling extensions.

III. SIMULATION STUDY

Prior Distributions and MCMC Sampling Diagnostics

For all the conditions of the simulation study, I specified covariates **x** as 0-1 dummy indicators to represent the 20 items. I assigned the following proper prior distributions on the model's parameters: $\theta_t \sim \text{normal}\left(0,\sigma_\theta^{\,2}\right)$ with $\sigma_\theta^{\,2} \sim \text{ig}\left(.1,.1\right)$ when the generating condition consisted of a unidimensional ability parameter, and $\boldsymbol{\theta}_t \sim \text{normal}_2\left(\mathbf{0},\Sigma_\theta\right)$ with $\Sigma_\theta \sim \text{iw}\left(4,.1\mathbf{I}_2\right)$ when the condition consisted of a two-dimensional ability parameter; $\boldsymbol{\tau}_h\left(j\right) \sim_{ind} \text{normal}_2\left(\boldsymbol{\mu}_j,.5\mathbf{I}_2\right)$, where $\boldsymbol{\mu}_j \sim \text{normal}_2\left(\mathbf{0},25\mathbf{I}_2\right)$; and $\upsilon_h\left(\mathbf{x}\right) \sim_{ind} \text{beta}\left(1,\alpha\right)$, where α was set to 2. $N_{\max}$ was set to 50.

During the analysis of each data set, I ran the MCMC sampling algorithm for 200,000 iterations in order to perform Bayesian posterior estimation. I discarded the first 100,000 samples (i.e., burn-in period) and saved every fifth sample thereafter for a total of 20,000 MCMC samples that were used for posterior inferences. According to standard convergence diagnostics (Geyer, 2011), the MCMC samples corresponding to model parameters displayed a good mixing of values to represent the posterior distribution for the parameters of interest.

The trace plots revealed that the parameter estimates stabilized after the burn-in period and the chains mixed well (i.e., the chain explores the support of the posterior distribution). Figure 1 contains the trace plots for the set of random step estimates for two different cases and items for the sample size condition of $N = 400$. The top two panels correspond to the step estimates for case 1, sampled during an analysis of the data generated under Condition 4 (i.e., unidimensional ability). The lower two panels correspond to the step estimates for case 148, sampled during an analysis of the data generated under Condition 9 (i.e., two-dimensional

ability). Figure 2 contains the trace plots for the set of random step estimates for two different persons and items for the sample size condition of $N = 800$. The top two panels correspond to the step estimates for case 543, sampled during an analysis of the data generated under Condition 4 (i.e., unidimensional ability). The bottom two panels correspond to the step estimates for case 629, sampled during an analysis of the data generated under Condition 9 (i.e., two-dimensional ability). In both figures, the trace plots correspond to DSF items.

Figure 3 and Figure 4 contain the trace plots of the saved samples of person ability estimate for two cases by two conditions (Condition 4 [unidimensional ability] and Condition 9 [two-dimensional ability]). The trace plots for all other estimated parameters were similar to those presented in these four figures.

I also examined the 95% Monte Carlo (MC) confidence intervals for the saved MCMC samples. They indicated that 20,000 saved MCMC samples led to sufficient precision of the posterior means of the person ability and random step estimates. Across the 10 generated data sets for the $N = 400$ condition, the 95% MC half-width confidence intervals for the person ability estimates had range (0.005, 0.016) and the random step estimates had range (0.007, 0.022). Across the 10 generated data sets for the $N = 800$ condition, the 95% MC half-width confidence intervals for the person ability estimates had range (0.005, 0.016) and the random step estimates had range (0.006, 0.023).

Predictive Performance of the Data

The predictive performance of the data provides a general overview of how the mDP version of the DDP-RM fared in relation to the comparison models. Table X contains the $D(m)$ values for each model by condition. In all sample size by data generating conditions, the mDP model outperformed all comparison models with respect to predictive performance of the data.

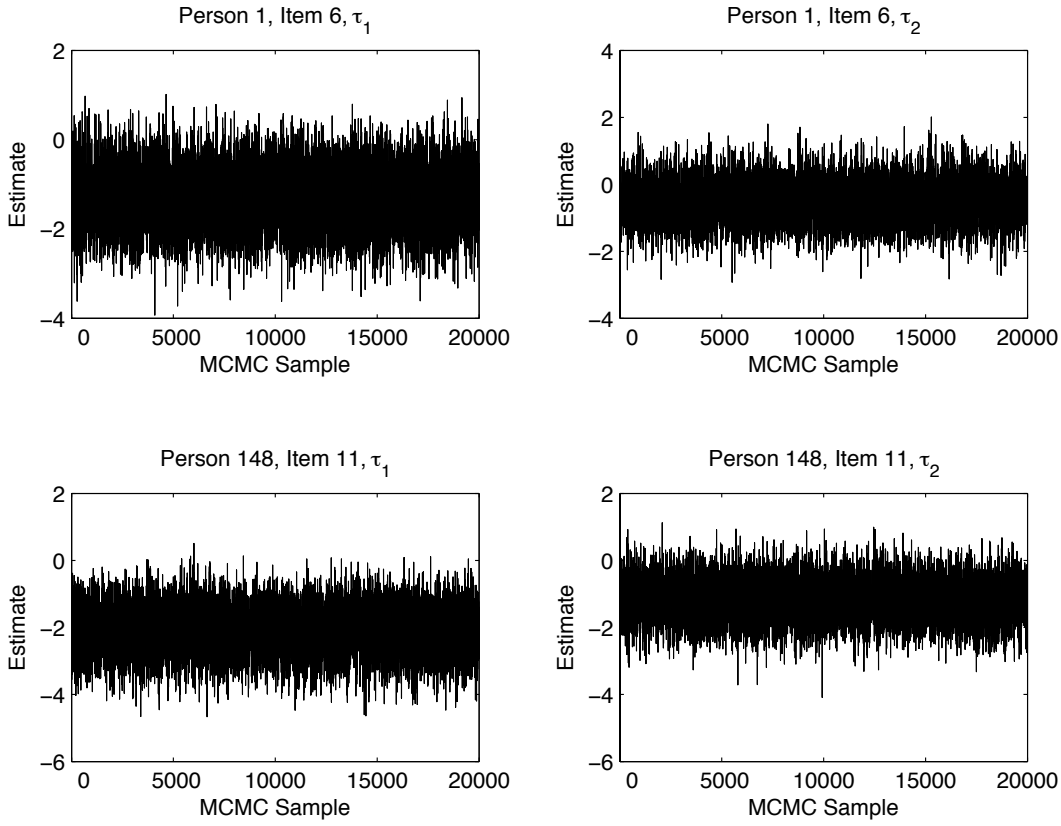

Figure 1. Trace plots of the set of random step estimates for two items when the sample size condition was $N = 400$. The trace plots in the top two panels correspond to the step estimates sampled during an analysis of the data generated under Condition 4 and the trace plots in the bottom two panels correspond to the step estimates sampled during an analysis of the data generated under Condition 9.

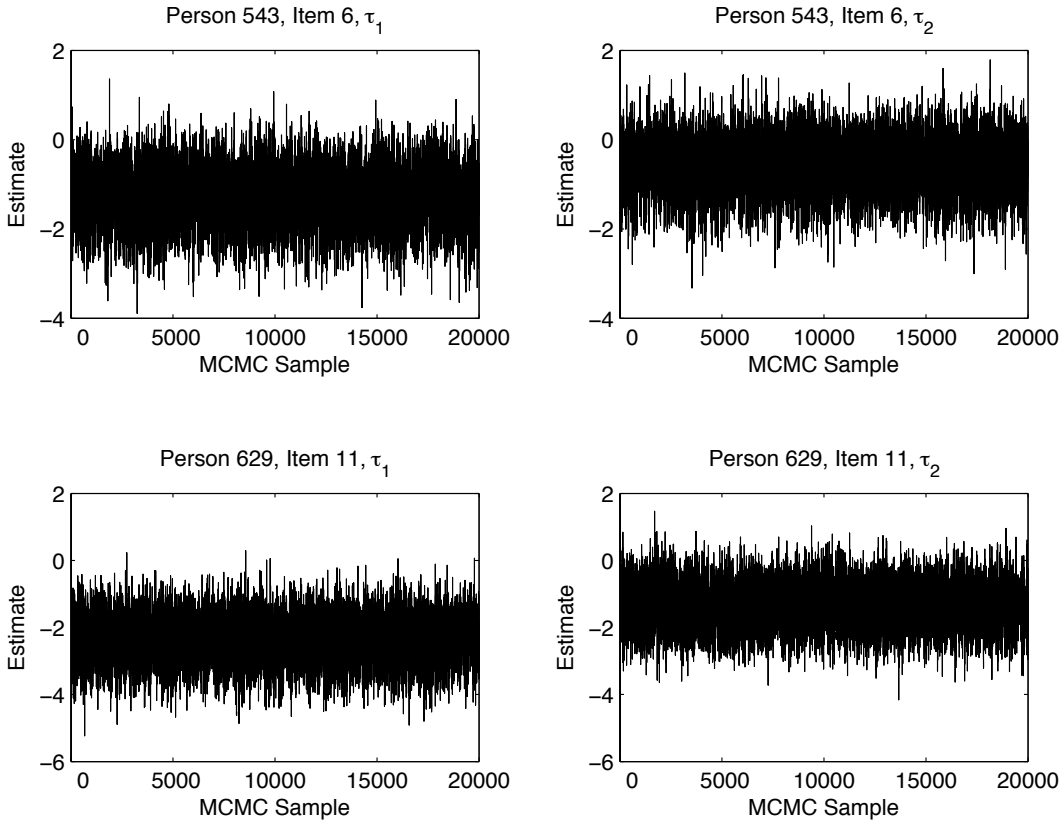

Figure 2. Trace plots of the set of random step estimates for two items when the sample size condition was $N = 800$. The trace plots in the top two panels correspond to the step estimates sampled during an analysis of the data generated under Condition 4 and the trace plots in the bottom two panels correspond to the step estimates sampled during an analysis of the data generated under Condition 9.

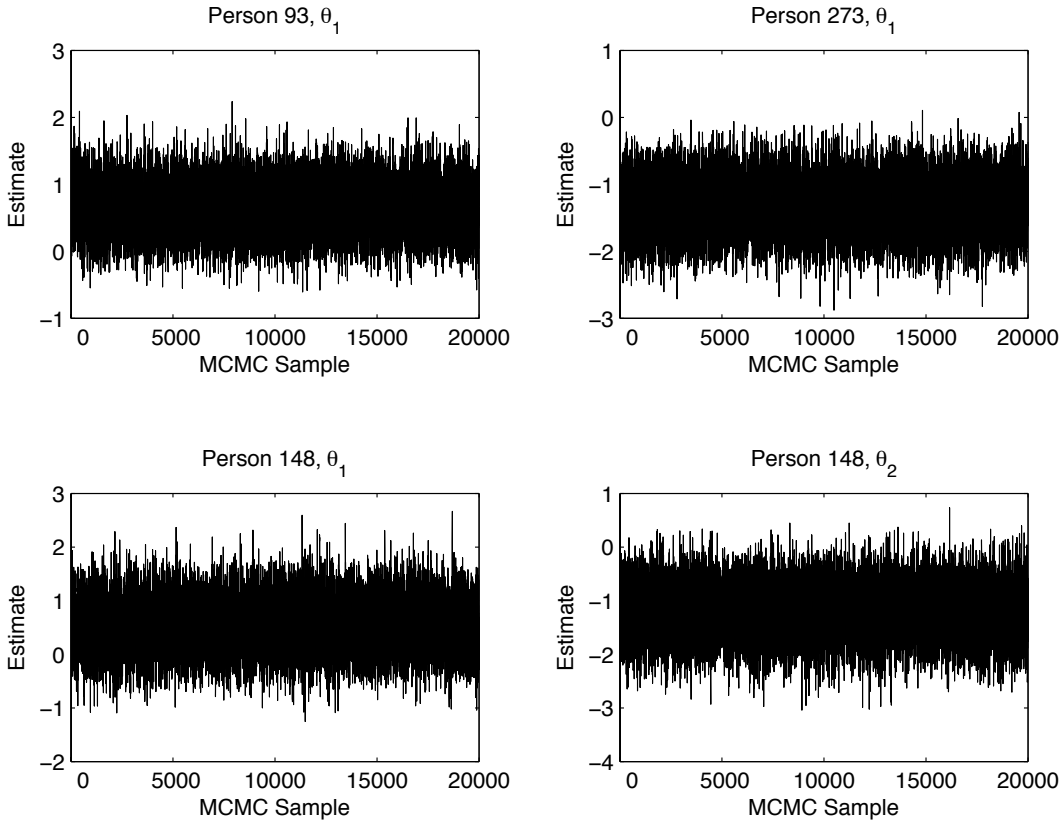

Figure 3. Trace plots of the ability estimate(s) for two cases when the sample size condition was $N = 400$. The trace plots in the top two panels correspond to the person ability estimates sampled during an analysis of the data generated under Condition 4 and the trace plots in the bottom two panels correspond to the person ability estimates sampled during an analysis of the data generated under Condition 9.

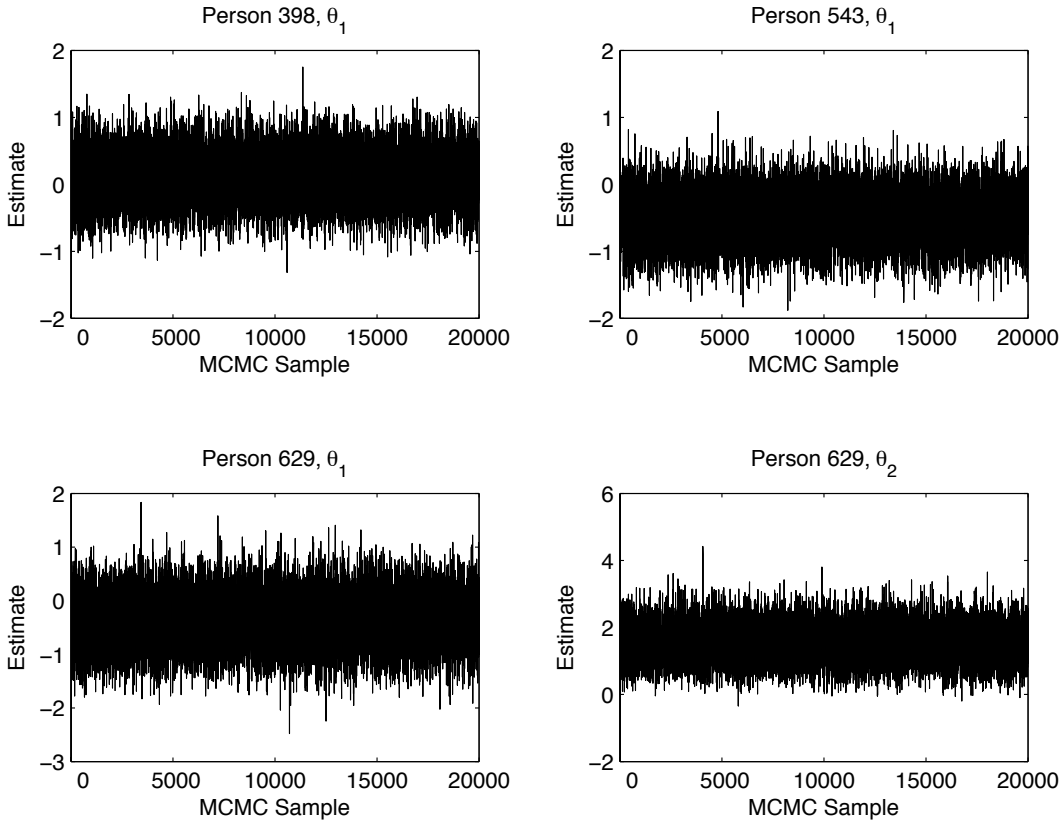

Figure 4. Trace plots of the ability estimate(s) for two cases when the sample size condition was $N = 800$. The trace plots in the top two panels correspond to the person ability estimates sampled during an analysis of the data generated under Condition 4 and the trace plots in the bottom two panels correspond to the person ability estimates sampled during an analysis of the data generated under Condition 9.

TABLE X

THE *D*(*m*) VALUES BY CONDITION AND MODEL

	Condition	mDP	Mixture Rasch[a]	GPCM	PCM	RSM	NRM	GRM
	1	5769	5877 (2)	6070	6090	6094	6057	6132
	2	5723	5928 (1)	6023	6027	6034	6012	6094
	3	5672	5702 (2)	5951	5955	5965	5944	6022
	4	5753	5837 (1)	6036	6045	6047	6035	6093
N = 400	5	5966	6165 (3)	6257	6268	6279	6250	6325
	6	5626		5832	5850	5851	5830	5911
	7	5558		5797	5803	5813	5780	5860
	8	5748		5983	5997	6001	6106	6053
	9	5632		5869	5872	5874	5868	5942
	10	5655		5863	5884	5885	5822	5950
	1	11317	11666 (1)	11841	11858	11862	11831	11971
	2	11387	11748 (1)	11928	11942	11956	11917	12052
	3	11214	11352 (2)	11805	11823	11842	11798	11930
	4	11712	12083 (2)	12304	12318	12326	12283	12445
N = 800	5	12070	12569 (5)	12734	12769	12784	12729	12873
	6	11307		11705	11714	11715	12518	11850
	7	11237		11639	11663	11678	11675	11783
	8	11302		11714	11726	11736	11751	11877
	9	11289		11670	11684	11687	11720	11788
	10	11250		11648	11650	11652	11702	11792

Note. mDP = multiple Dirichlet process model, GPCM = generalized partial credit model, PCM = partial credit model, RSM = rating scale model, NRM = nominal response model, and GRM = graded response model. Lower value indicates better predictive performance (i.e., better model fit to the data).

[a]Due to limitations of the Winmira software, the finite-mixture Rasch model was used to analyze only the data in which the person ability was unidimensional (i.e., Conditions 1-5). The number of estimated latent classes is in the parentheses next to its *D*(*m*) value. For each condition, the optimal number of latent classes for the finite-mixture PCM was identified using the AIC.

Under both sample size conditions, for the first five data generating conditions in which the finite-mixture Rasch model was one of the comparison models, it displayed far better predictive performance than the other comparison IRT models. Within the $N = 400$ condition, the finite-mixture PCM indicated that two latent classes were present in the data set corresponding to Conditions 1 and 3 when these conditions consisted of only a single latent class. Moreover, the finite-mixture PCM indicated that one and three latent classes were present in the data corresponding to Conditions 4 and 5, respectively, but the data within these two conditions were generated to reflect two latent classes. Within the $N = 800$ condition, the finite-mixture PCM incorrectly identified 2 and 5 latent classes as being present in the data corresponding to Conditions 3 and 5, respectively. The data in Conditions 3 and 5 were generated to reflect one and two latent classes, respectively. Within this sample size condition, the finite-mixture PCM correctly identified that two latent classes were present in the data for Condition 4.

Even though the finite-mixture PCM outperformed the other comparison IRT models, for the $N = 400$ condition, the mDP outperformed the finite-mixture PCM by at least 30 $D(m)$ units and bested the finite-mixture PCM by 84 and 199 $D(m)$ units for the two different unidimensional DSF conditions (i.e., Conditions 4 and 5). For the $N = 800$ condition, the mDP model outperformed the finite-mixture PCM by at least 138 $D(m)$ units and outperformed it by 371 and 499 $D(m)$ units for the two different unidimensional DSF conditions (i.e., Conditions 4 and 5). The mDP model's ability to detect the appropriate number of latent classes for the DSF conditions is discussed in the section in which the posterior mean estimates of the mixing distribution are presented.

Root Mean Squared Deviation

The root mean squared deviation (RMSD) provides a means to examine how well the mDP model and the comparison models recovered the generating values used for the person ability and item category step parameters. Recall that values closer to 0 indicate better recovery.

RMSD for the Item Category Steps

Table XI contains the RMSD values that indicate how well the models recovered the generating step values for each condition. Recall that the RMSD involving the category steps was only calculated for the models in which the link functions were defined in terms of adjacent categories, which were the mDP model, finite-mixture PCM, GPCM, PCM, and the RSM**.** Regardless of the sample size condition, the PCM or RSM displayed the best recovery of the generating values within each data generating condition. This finding is not surprising given that the generating conditions relied on the RMS, PCM, or a blocked RSM. The benefit of such a match between the generating model and the model used to estimate the category step parameters was apparent in Conditions 1 and 6. In Condition 1, the unidimensional RSM was the generating model, and in Condition 6, a two-dimensional RSM was the generating model. Indeed, the RMSD was the lowest for the RMS for these two conditions. When the sample size condition was $N = 400$, the RSM had values 0.081 and 0.095 for Conditions 1 and 6, respectively, and when the sample size condition was $N = 800$, the RSM had values 0.063 and 0.043 for Conditions 1 and 6, respectively.

Unfortunately, the infinite-mixture model (mDP model) and the finite-mixture PCM did not perform as well as the PCM and RSM, even for the conditions in which the data for a subset of items were generated to have DSF or DIF (i.e., Conditions 4, 5, 9, and 10). Overall, however, the mDP model outperformed the finite-mixture PCM. In fact, for Condition 5, the finite-mixture

TABLE XI

THE ROOT MEAN SQUARED DEVIATION BETWEEN CATEGORY STEP ESTIMATES AND THE CATEGORY STEP GENERATING VALUES BY SAMPLE SIZE CONDITION AND MODEL

	Condition	mDP	Mixture PCM	GPCM	PCM	RSM
	1	0.218	0.696	0.175	0.171	0.081
	2	0.335	0.285	0.304	0.288	0.225
	3	0.199	0.471	0.128	0.128	0.206
	4	0.278	0.272	0.213	0.216	0.195
N=400	5	0.290	1.052	0.250	0.240	0.280
	6	0.196		0.172	0.131	0.095
	7	0.362		0.302	0.305	0.226
	8	0.195		0.151	0.115	0.158
	9	0.314		0.286	0.276	0.244
	10	0.238		0.215	0.197	0.183
	1	0.185	0.081	0.091	0.088	0.063
	2	0.304	0.252	0.248	0.250	0.214
	3	0.198	0.428	0.137	0.099	0.194
	4	0.240	0.638	0.190	0.175	0.172
N=800	5	0.277	1.223	0.230	0.229	0.279
	6	0.158		0.087	0.080	0.043
	7	0.294		0.245	0.220	0.181
	8	0.197		0.114	0.099	0.143
	9	0.282		0.235	0.223	0.203
	10	0.232		0.177	0.165	0.147

Note. mDP = multiple Dirichlet process model, GPCM = generalized partial credit model, PCM = partial credit model, and RSM = rating scale model. Lower value indicates better recovery of generating values. Due to limitations of the Winmira software, the finite-mixture Rasch model was used to analyze only the data in which the person ability was unidimensional (i.e., Conditions 1-5). The number of estimated latent classes is in the parentheses next to its *D*(*m*) value. For each condition, the optimal number of latent classes for the finite-mixture PCM was identified using the AIC.

PCM displayed very poor recovery, with the RMSD for this condition being at least three times larger than the next largest RMSD when the sample size was $N = 400$ and at least four times larger than the next largest RMSD when the sample size was $N = 800$.

RMSD for the Person Abilities (θ)

Table XII contains the RMSD values that indicate how well the models recovered the generating ability values for each condition. With respect to the recovery of the generating values, while overall the PCM and RMS had the lowest RMSD values, they did not display the same superior performance over the mDP model as observed in the recovery of the category step values. Among the models, the finite-mixture PCM performed slightly worse than the other models in terms of recovering the generating ability values. The mDP model and the finite-mixture PCM did not display an advantage over the other models in terms of recovering the generating ability values, even for the conditions in which the data for a subset of items were generated to reflect DSF or DIF between two groups (i.e., Conditions 4, 5, 9, and 10).

Posterior Mean Estimates of the Mixing Distribution

The posterior mean estimates of the mixing distributions, $G_{\mathbf{x}}(\tau_1)$ and $G_{\mathbf{x}}(\tau_2)$, have been shown to indicate whether an item has DSF (Fujimoto & Karabatsos, 2014). Thus, looking at posterior means and standard deviations (*SD*) of an item's category steps could be the first indication of whether an item is problematic. Then examining the marginal posterior predictive densities of the mixing distribution could be beneficial. In this section, I review the posterior means and standard deviations for the analysis of the data generated under Conditions 3, 4, and 10 for both sample size conditions. The analysis of the data generated under other conditions displayed similar patterns as those presented in this section.

TABLE XII

THE ROOT MEAN SQUARED DEVIATION BETWEEN THE ABILITY (THETA) ESTIMATES AND THE ABILITY GENERATING VALUES BY SAMPLE SIZE CONDITION AND MODEL

	Condition	mDP	Mixture PCM	GPCM	PCM	RSM	NRM	GRM
	1	0.365	0.402	0.361	0.358	0.358	0.360	0.368
	2	0.349	0.379	0.344	0.341	0.341	0.344	0.350
	3	0.362	0.401	0.358	0.356	0.356	0.361	0.360
	4	0.352	0.412	0.347	0.345	0.345	0.350	0.348
N=400	5	0.355	0.465	0.350	0.351	0.351	0.350	0.355
	6	0.460		0.457	0.455	0.455	0.459	0.461
	7	0.451		0.454	0.454	0.453	0.455	0.459
	8	0.451		0.455	0.451	0.451	0.459	0.456
	9	0.496		0.494	0.490	0.490	0.496	0.496
	10	0.454		0.455	0.452	0.452	0.455	0.460
	1	0.355	0.378	0.349	0.346	0.346	0.350	0.354
	2	0.345	0.369	0.335	0.335	0.335	0.337	0.339
	3	0.370	0.435	0.357	0.354	0.354	0.355	0.360
	4	0.348	0.474	0.344	0.344	0.344	0.346	0.346
N=800	5	0.347	0.468	0.346	0.345	0.344	0.347	0.346
	6	0.461		0.458	0.457	0.457	0.460	0.460
	7	0.456		0.454	0.452	0.452	0.455	0.459
	8	0.464		0.459	0.457	0.457	0.460	0.463
	9	0.467		0.464	0.462	0.462	0.466	0.466
	10	0.466		0.464	0.462	0.462	0.465	0.468

Note. mDP = multiple Dirichlet process model, GPCM = generalized partial credit model, PCM = partial credit model, RSM = rating scale model, NRM = nominal response model, and GRM = graded response model. Lower value indicates better recovery of generating values. Due to limitations of the Winmira software, the finite-mixture Rasch model was used to analyze only the data in which the person ability was unidimensional (i.e., Conditions 1-5). The number of estimated latent classes is in the parentheses next to its $D(m)$ value. For each condition, the optimal number of latent classes for the finite-mixture PCMs was identified using the AIC.

Condition 3

Recall that in this condition, all items were free of DSF. Thus, the posterior *SD* estimates should be similar across all items and thresholds, regardless of the sample size condition. Table XIII contains the item category step posterior means and *SD* estimates for this data generating condition, presented for each sample size condition.

Sample size condition of 400. For this sample size condition, the posterior *SD* estimates of the category steps had range (0.50, 0.66). The posterior *SD* estimates were fairly similar across all items and steps, as expected given that the data for all items were generated to be free of DSF. The posterior mean estimates follow the general pattern of difficulty as the generating item ordering for this condition. For the generating item values please see Table III. The easier items (i.e. items with lower category step posterior mean estimates) were the first few items, and the more difficult items (i.e., items with higher category step posterior mean estimates) were the last few items.

Figure 5 includes the marginal posterior density estimates of the category steps for Items 5, 12, and 17. The densities for these three items had similar forms. That is, they were unimodal and the variability did not drastically differ across these three items, as expected given that in this condition, the data for all items were generated to lack DSF. Similar conclusions were drawn on the remaining items within this condition for which densities were not presented.

Sample size condition of 800. For this sample size condition, the posterior *SD* estimates of the category steps had range (0.43, 1.04). The *SD* value of 1.04 corresponded to Item 5's second step. Aside from this one value, the next largest posterior *SD* estimate was 0.70, which was for Item 13's first step. Given that the data for all items within this condition were generated to be free of DSF, the large *SD* observed for Item 5 could be viewed as an anomaly because the

TABLE XIII

POSTERIOR MEAN AND STANDARD DEVIATION (IN PARENTHESES) ESTIMATES OF THE ITEM CATEGORY STEPS BY SAMPLE SIZE FOR CONDITION 3

	N = 400		*N* = 800	
Item	Step 1	Step 2	Step 1	Step 2
1	–2.24 (0.60)	–1.33 (0.52)	–2.03 (0.62)	–1.14 (0.68)
2	–1.81 (0.58)	–0.96 (0.51)	–1.96 (0.60)	–0.81 (0.59)
3	–2.23 (0.61)	–0.51 (0.66)	–2.11 (0.63)	–0.28 (0.55)
4	–1.71 (0.63)	–0.81 (0.61)	–1.33 (0.59)	–0.80 (0.56)
5	–1.43 (0.66)	–0.58 (0.54)	–1.45 (0.60)	–0.41 (1.04)
6	–1.83 (0.60)	0.05 (0.59)	–1.47 (0.63)	0.12 (0.62)
7	–0.74 (0.57)	–0.42 (0.54)	–1.02 (0.56)	–0.24 (0.53)
8	–1.02 (0.60)	0.17 (0.55)	–1.00 (0.52)	0.04 (0.63)
9	–0.94 (0.57)	0.53 (0.54)	–0.84 (0.51)	0.57 (0.60)
10	–0.45 (0.61)	0.08 (0.61)	–0.31 (0.50)	0.24 (0.54)
11	–0.46 (0.66)	0.26 (0.60)	–0.44 (0.48)	0.66 (0.51)
12	–0.64 (0.57)	1.17 (0.64)	–0.48 (0.57)	1.12 (0.62)
13	0.00 (0.50)	0.95 (0.62)	0.21 (0.70)	0.71 (0.55)
14	0.16 (0.55)	1.29 (0.57)	0.01 (0.51)	1.35 (0.60)
15	–0.07 (0.58)	1.59 (0.60)	–0.15 (0.43)	1.97 (0.54)
16	0.93 (0.65)	1.31 (0.62)	0.53 (0.58)	1.35 (0.56)
17	0.47 (0.58)	1.64 (0.58)	0.59 (0.62)	1.87 (0.63)
18	0.49 (0.56)	1.86 (0.57)	0.63 (0.65)	2.04 (0.58)
19	1.24 (0.59)	1.66 (0.60)	1.23 (0.47)	1.95 (0.55)
20	1.00 (0.58)	1.99 (0.59)	1.12 (0.46)	2.22 (0.60)

Note. All items in this condition were free of DSF.

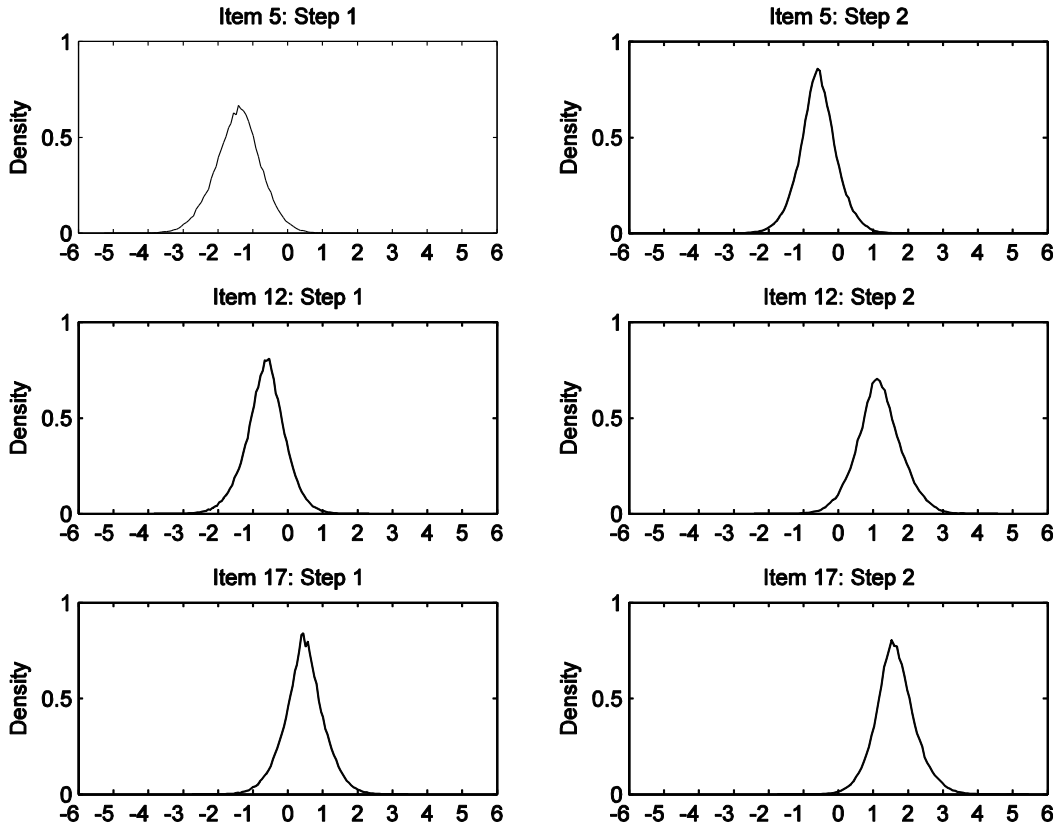

Figure 5. Posterior predictive density estimates of the category steps for Items 5, 12, and 17 for Condition 3 ($N = 400$).

posterior *SD* estimates corresponding to all other items had fairly similar values, as expected under this generating condition.

The posterior mean estimates followed the general pattern of difficulty in item ordering as the generating items. For the generating values, please refer to Table III. The easier items were the first few items and the items progressively became more difficult, as indicated by the increase in estimated values of the items' step posterior means.

Figure 6 includes the marginal posterior density estimates for the category steps corresponding to Items 5, 12, and 17. The density for Item 5's second step had a bimodal form. This bimodal form was why the posterior *SD* estimate of this step was noticeably larger than the estimates of the category steps for all other items. The bimodality was most likely a result of the random data generating process because, under this condition, the data for all items were intended to be generated to be free of DSF. The other densities in Figure 6 were unimodal and had similar variances. The posterior densities for the remaining items not displayed in this figure were also unimodal with small variability as the densities for Items 12 and 17.

Condition 4

In this condition, the data corresponding to Items 3, 6, 9, 12, 15, and 18 were generated to have a 1-logit DSF in the second category steps. Thus, the posterior *SD* estimates corresponding to the second steps for these items should be larger than the posterior *SD* estimates for the other items (i.e., DSF-free items). Table XIV contains the item category step posterior means and *SD* estimates for this data generating condition, presented for each sample size condition.

Sample size condition of 400. For this sample size condition, the posterior *SD* estimates of the category steps had range (.51, .75). The items with larger estimates did not correspond to

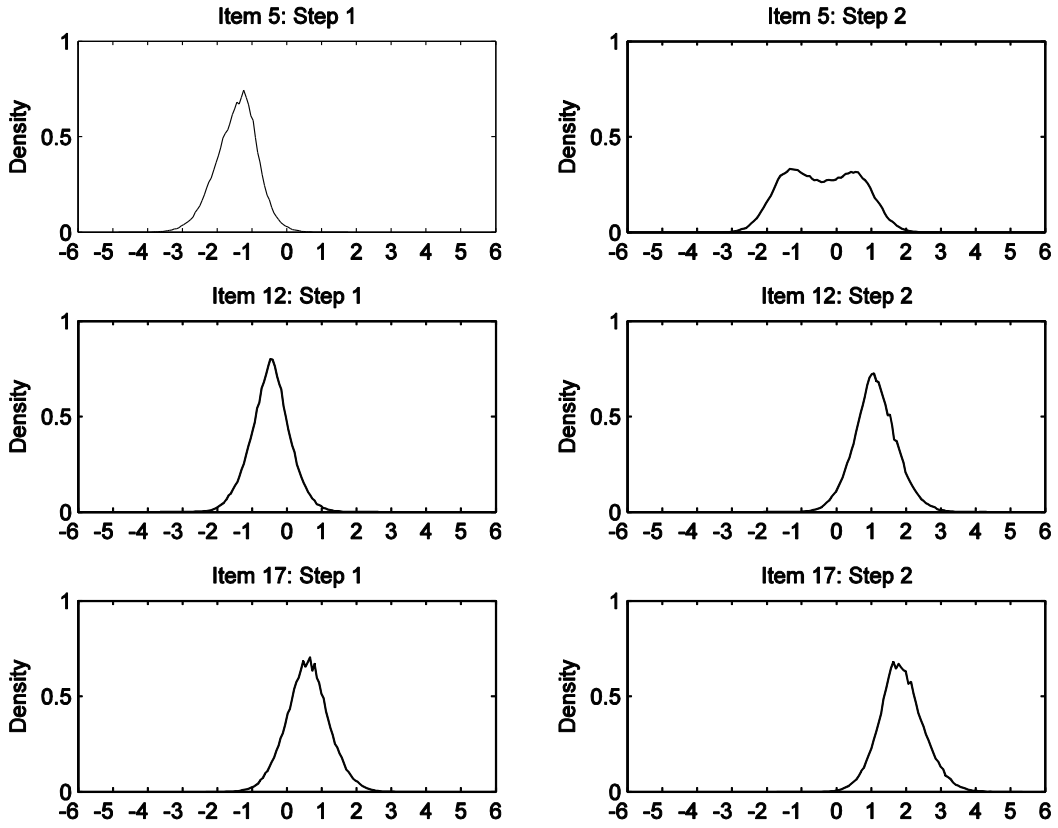

Figure 6. Posterior predictive density estimates of the category steps for Items 5, 12, and 17 for Condition 3 ($N = 800$).

TABLE XIV

POSTERIOR MEAN AND STANDARD DEVIATION (IN PARENTHESES) ESTIMATES OF THE ITEM CATEGORY STEPS BY SAMPLE SIZE FOR CONDITION 4

	N = 400		*N* = 800	
Item	Step 1	Step 2	Step 1	Step 2
1	–2.42 (0.62)	–1.10 (0.63)	–2.38 (0.59)	–0.98 (0.60)
2	–2.16 (0.64)	–1.28 (0.58)	–2.14 (0.63)	–1.24 (0.64)
3	–1.77 (0.67)	–0.92 (0.58)	–1.63 (0.57)	–0.98 (0.74)
4	–1.30 (0.62)	–0.77 (0.53)	–1.60 (0.61)	–0.53 (0.49)
5	–1.62 (0.61)	–0.87 (0.62)	–1.48 (0.55)	–0.51 (0.68)
6	–1.47 (0.61)	–0.55 (0.51)	–1.22 (0.56)	–0.40 (0.50)
7	–1.37 (0.66)	–0.25 (0.55)	–1.18 (0.59)	–0.04 (0.63)
8	–1.04 (0.56)	–0.38 (0.52)	–0.95 (0.54)	–0.12 (0.64)
9	–0.95 (0.56)	0.36 (0.56)	–0.63 (0.72)	0.03 (0.47)
10	–0.47 (0.52)	0.53 (0.53)	–0.61 (0.48)	0.56 (0.61)
11	–0.59 (0.61)	0.36 (0.59)	–0.64 (0.63)	0.57 (0.56)
12	–0.47 (0.59)	0.45 (0.59)	–0.05 (0.56)	0.44 (0.47)
13	–0.02 (0.54)	1.02 (0.75)	–0.13 (0.51)	1.19 (0.65)
14	–0.30 (0.57)	0.94 (0.66)	0.18 (0.47)	0.96 (0.53)
15	0.06 (0.64)	0.91 (0.59)	0.30 (0.58)	0.93 (0.55)
16	0.25 (0.52)	1.49 (0.57)	0.47 (0.64)	1.24 (0.63)
17	0.68 (0.54)	1.44 (0.60)	0.62 (0.62)	1.49 (0.52)
18	0.79 (0.59)	1.54 (0.63)	0.59 (0.60)	1.53 (0.64)
19	1.04 (0.53)	1.73 (0.57)	0.86 (0.44)	1.99 (0.60)
20	1.00 (0.57)	2.05 (0.58)	0.97 (0.55)	1.82 (0.54)

Note. For this condition, Items 3, 6, 9, 12, 15, and 18 had 1-logit DSF in the upper steps between two groups.

the DSF items. For instance, the estimate for Item 13's second step was 0.75, but the data for this item was generated to be free of DSF. Item 6, which was a DSF item, had an estimate of 0.51 for the second step. Thus, the size of the posterior *SD* estimates did not provide insight into which items had DSF given the simulation study conditions within this chapter.

The posterior mean estimates of the category steps followed the general pattern of difficulty in item ordering as the generating condition. For the generating values, please see Table IV. The easier items were the first few items and the items progressively become more difficult, as indicated by the increase in values of the posterior mean estimates for the steps. For the items that had DSF, the posterior mean estimates of the category steps were approximately the average of the generating step values for the two groups. For example, the generating values for Item 6's second step were 0.04 and –0.96 for Groups 1 and 2, respectively. The posterior mean estimate of –0.55 for this step was close to the average of the generating value of –0.46 for this step. The DSF was only in the second category step. Thus, both groups had the same generating values for the first step for this item.

Figure 7 contains the marginal posterior density estimates of the category steps for Items 6, 12, and 13. The densities for Items 6 and 12 had similar forms, which were unimodal and had similar variability. The densities for Item 13 were also unimodal, but the variability in the density for the second step was greater, which reflects the larger posterior *SD* estimate for this step that was previously noted. The densities for the other items not presented in this figure were also unimodal and did not drastically differ in terms of variability in the densities.

Sample size condition of 800. For this sample size condition, the posterior *SD* estimates of the category steps had range (0.44, 0.74). The largest estimate was for the second step for Item 3. This item was a DSF item. However, the value of 0.74 was only slightly larger than the

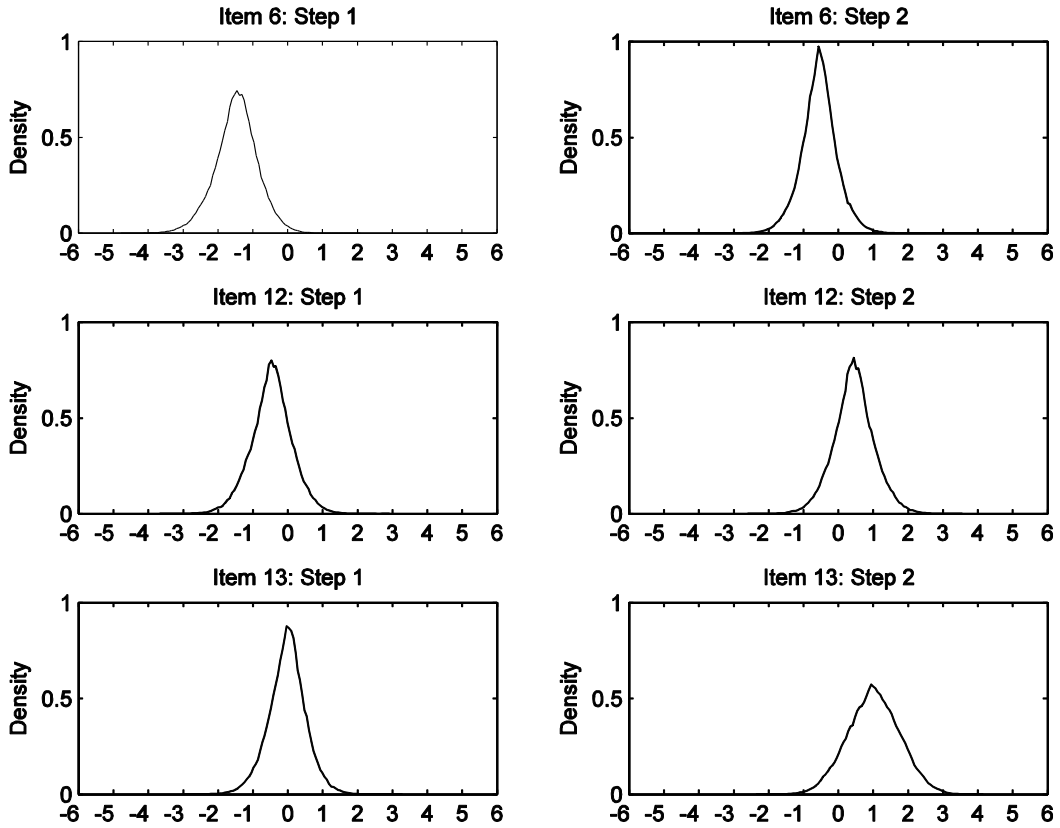

Figure 7. Posterior predictive density estimates of the category steps for Items 6, 12, and 13 for Condition 4 ($N = 400$).

average posterior *SD* estimates of the steps for all other items, including DSF and DSF-free items. Item 9, which was another DSF item, had a posterior *SD* estimate of 0.72 and 0.47 for the first and second category steps, respectively. Unfortunately, the size of the estimates should be reversed for this item because the data for this item were generated only to have DSF in the second step. The posterior *SD* estimates of the remaining category steps fell between 0.44 and 0.68. Thus, even with a sample size of 800, the posterior *SD* estimates did not clearly identify the DSF items given the simulation conditions.

The posterior mean estimates for the category steps followed the general pattern of difficulty in item ordering as the generating items. For the generating step values for this condition, please see Table IV. The easier items were the first few items and the items progressively became more difficult, as indicated by the increase in values of the posterior mean estimates for the steps. For the items that had DSF, the posterior mean estimates of the category steps were roughly the average of the generating step values for the two groups. The posterior mean estimate of the second category step for Item 6 was –0.40, which was close to the average of the generating value of –0.46 for this step.

Figure 8 includes the marginal posterior density estimates of the category step distributions for Items 3, 9, and 11. All of the densities had unimodal form, but the density for Item 3's second step and Item 9's first step were slightly wider and flatter, which reflects the larger posterior *SD* estimates previously noted for these item. The forms of the density for the second category steps for Item 9 and the densities for both steps for Item 11 were fairly similar even though the DSF in Item 9 should appear in the density of the second step and Item 11 was a DSF-free item. The posterior densities for the items not presented in this figure were also unimodal and had similar variability as the densities for Items 9 and 11. Even with a sample size

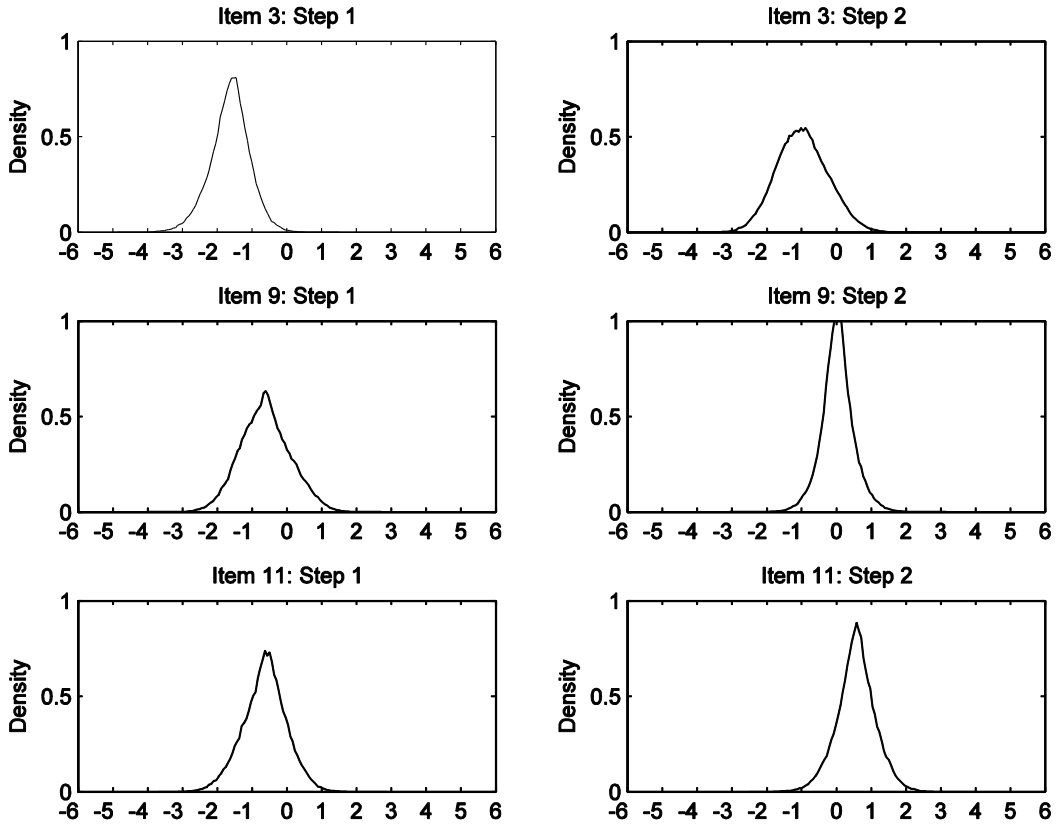

Figure 8. Posterior predictive density estimates of the category steps for Items 3, 9, and 11 for Condition 4 ($N = 800$).

of 800, the posterior densities did not provide clear indications of which items had DSF given the simulation conditions for this study.

Condition 10

For this condition, the generating ability distribution was two-dimensional, with the first 10 items being indicators for the first dimension of the latent trait and the second 10 items being indictors for the second dimension of the latent trait. Additionally, the data for Items 11, 13, 17, and 19 were generated to have greater DSF effect in the second steps (1-logit difference in the generating parameters for the two groups) between the two groups than the first steps (.5-logit difference in the generating parameters for the two groups). The data for Items 14 and 15 were generated to reflect DIF between two the groups (.5-logit difference in the generating parameters for the two groups for steps 1 and 2). Because these were the DIF or DSF items, the posterior *SD* estimates corresponding to these items were expected to be larger than the estimates for the other items. Table XV contains the item category step posterior mean and *SD* estimates for this data condition, presented for each sample size condition.

Sample size condition of 400. Across the 20 items, the posterior *SD* estimates for the item category steps had range (0.50, 0.76). The second step for Item 1 had the largest estimate (posterior *SD* = 0.76), which was a DSF-free item. Among the DSF and DIF items, the posterior *SD* estimates had range (0.52, 0.64). These values fell within a very narrow range (0.12 difference between the minimum and maximum within this subset of items). Moreover, these values were highly similar to the estimates for the items free of DSF and DIF. Thus, the size of the posterior *SD* estimates did not appear to provide insight into which items had DSF given the generating conditions.

TABLE XV

POSTERIOR MEAN AND STANDARD DEVIATION (IN PARENTHESES) ESTIMATES OF THE ITEM CATEGORY STEPS BY SAMPLE SIZE FOR CONDITION 10

	$N = 400$		$N = 800$	
Item	Step 1	Step 2	Step 1	Step 2
1	–2.09 (0.66)	–1.04 (0.76)	–2.12 (0.60)	–0.97 (0.48)
2	–1.61 (0.60)	–0.93 (0.64)	–1.69 (0.59)	–0.57 (0.57)
3	–1.11 (0.59)	–0.36 (0.56)	–1.07 (0.53)	–0.16 (0.58)
4	–0.89 (0.50)	0.16 (0.58)	–0.82 (0.61)	0.16 (0.71)
5	–0.46 (0.54)	0.31 (0.56)	–0.72 (0.54)	0.52 (0.57)
6	–0.33 (0.57)	0.81 (0.56)	–0.22 (0.50)	0.95 (0.52)
7	0.03 (0.58)	1.31 (0.59)	0.15 (0.54)	1.12 (0.58)
8	0.53 (0.56)	1.37 (0.62)	0.47 (0.62)	1.56 (0.54)
9	0.84 (0.60)	1.69 (0.63)	0.89 (0.62)	2.01 (0.63)
10	1.13 (0.55)	2.34 (0.62)	1.30 (0.58)	2.27 (0.57)
11	–2.59 (0.64)	–1.34 (0.56)	–2.46 (0.61)	–1.53 (0.56)
12	–1.70 (0.62)	–1.12 (0.65)	–1.73 (0.58)	–1.28 (0.55)
13	–1.63 (0.56)	–0.80 (0.52)	–2.14 (0.62)	–0.80 (0.50)
14	–1.60 (0.63)	–0.67 (0.61)	–1.75 (0.62)	–0.77 (0.54)
15	–1.30 (0.57)	–0.37 (0.57)	–1.11 (0.55)	–0.32 (0.55)
16	–0.51 (0.64)	0.38 (0.61)	–0.53 (0.64)	0.24 (0.61)
17	–0.45 (0.58)	0.51 (0.54)	–0.47 (0.53)	0.68 (0.56)
18	0.06 (0.66)	1.04 (0.57)	0.12 (0.51)	0.96 (0.67)
19	0.41 (0.63)	1.39 (0.56)	0.17 (0.50)	1.29 (0.64)
20	1.01 (0.55)	1.77 (0.61)	0.78 (0.58)	1.83 (0.63)

Note. The items were indicators for two dimensions of the latent trait. Items 1 through 10 were indicators for the first ability dimension, and Items 11 through 20 were indicators for the second ability dimension. Items 11, 13, 17, and 19 were generated to have 1-logit DSF in the upper step and .5-logit DSF in the lower step. Items 14 and 15 had .5-logit DIF (i.e., both steps had .5-logit difference in the generating values between the two groups).

The posterior mean estimates followed the general pattern of difficulty ordering as the generating items. For the generating step values for this condition, please see Table VIII. Within the subset of items associated with the first ability dimension (i.e., Items 1 through 10), the first few items within this subset had lower posterior mean estimates, indicating that these items were easier, and the last few items within this subset had higher posterior mean estimates, indicating that these items were more difficult within this subset. Likewise, within the subset of items associated with the second ability dimension (i.e., Items 11 through 20), the first few items within this subset had lower posterior means and the last few items within this subset had higher posterior mean values. For the items that had DSF or DIF, the posterior mean estimates of the steps were roughly the average of the generating step values for the two groups. For example, Group 1's generating values for Item 11's category steps were –2.20 and –0.95 and Group 2's generating values for this same item were –2.70 and –1.95. The posterior mean estimates for the first and second step for this item were –2.59 and –1.35, respectively, which were close to the average of the generating values for the two groups of –2.45 and –1.45, respectively.

Figure 9 contains the posterior density plots for the category steps corresponding to Items 11, 12, and 17. The posterior distributions for the three items were unimodal and had similar variability. Yet, Items 11 and 17 were the DSF items and Item 12 was a DSF-free item. The posterior densities for other items not included in this figure had similar forms. Thus, the posterior densities did not appear to clearly indicate which items had DSF given the simulation conditions for this study.

Sample size condition of 800. Across the 20 items, the posterior *SD* estimates for the item category steps had range (0.48, 0.71). The estimate of 0.71 corresponds to the second category step for Item 4, which is a DSF-free item. The estimates for the other items, regardless

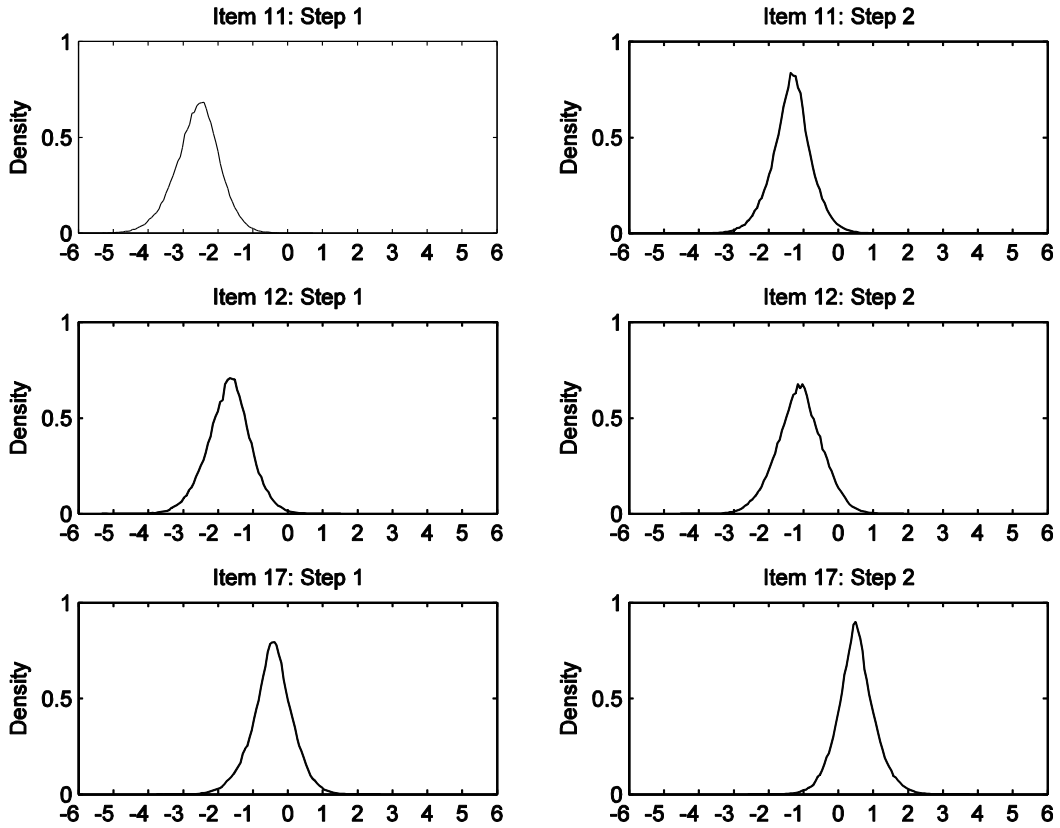

Figure 9. Posterior predictive density estimates of the category steps for Items 11, 12, and 17 for Condition 10 ($N = 400$).

of whether they were DSF or DIF items, were fairly similar. Thus, the size of the posterior *SD* estimates did not provide insight into which items had DSF given the generating conditions in this chapter.

The posterior mean estimates followed the general pattern of difficulty ordering as the generating items. For the generating step values for this condition, please see Table VIII. Within the subset of items associated with the first ability dimension (i.e., Items 1 through 10), the first few items within this subset had lower posterior mean estimates, indicating that these items were easier, and the last few items within this subset had higher estimates, indicating that these items were more difficult within this subset. Likewise, within the subset of items associated with the second ability dimension (i.e., Items 11 through 20), the first few items within this subset had lower posterior mean estimates and the last few items within this subset had higher estimates. For the DSF or DIF items, the posterior mean estimates of the steps were roughly the average of the generating step values for the two groups. For example, the posterior mean estimates for Item 11's first and second steps were –2.46 and –1.53, respectively, which were close to this item's average of the generating values for the two groups of –2.45 and –1.45, respectively.

Figure 10 includes the posterior density plots for the category steps corresponding to Items 4, 11, and 16. The posterior density plots for all three items were unimodal, with the density plot for Item 4's second step slightly wider and flatter (i.e., had slightly more variability). Yet, Items 4 and 16 were the DSF-free items and Item 11 was the DSF item. The posterior densities for the other items had similar forms as those in this figure. Thus, the posterior density plots did not appear to clearly differentiate the items that had DSF given the simulation conditions for this chapter.

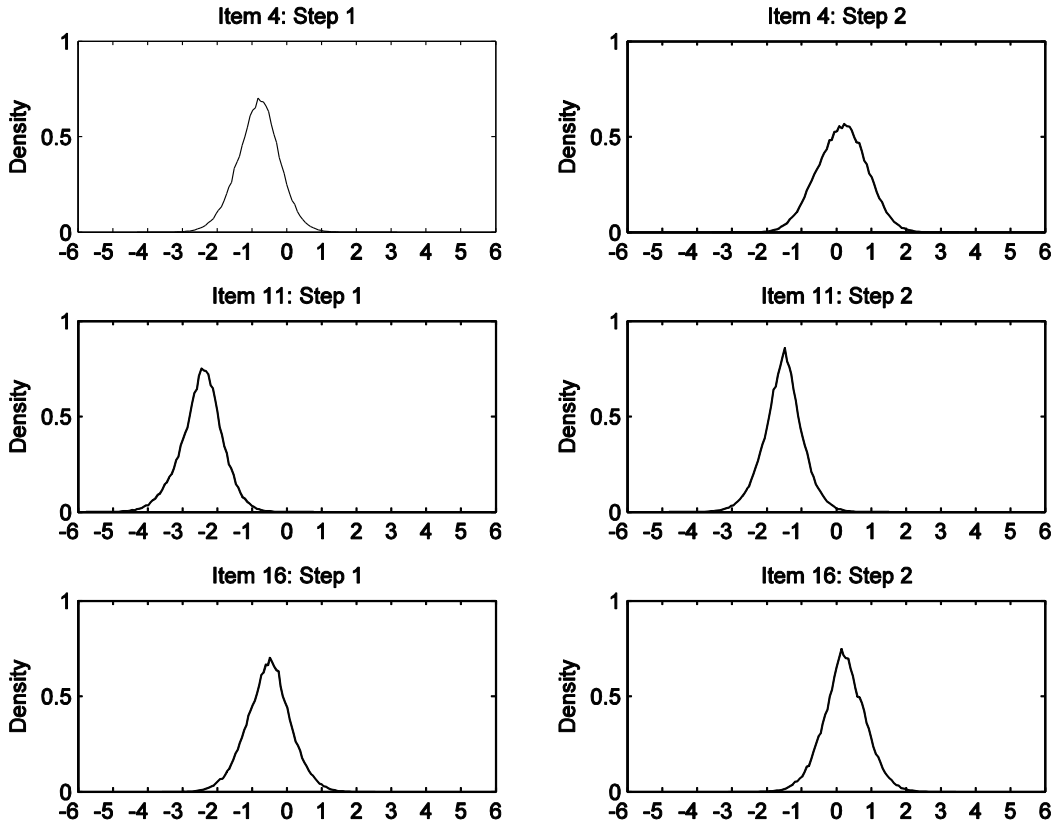

Figure 10. Posterior predictive density estimates of the category steps for Items 4, 11, and 16 for Condition 10 ($N = 800$).

IV. ANALYSIS OF REAL-LIFE DATA SETS

For this chapter, I analyzed the Verbal Aggression and the Acculturative Family Distancing (AFD) data sets. I first report the results of the analysis of the Verbal Aggression data and then the AFD data. For the analysis of these two data sets, the same proper priors were assigned on the parameters of the mDP model as those specified in the simulation study reported in the previous chapter. That is, $\boldsymbol{\theta}_t \sim \text{normal}_2\left(\mathbf{0}, \Sigma_\theta\right)$ with $\Sigma_\theta \sim \text{iw}\left(4, .1\mathbf{I}_2\right)$ when the condition called for a two-dimensional ability parameter (for the Verbal Aggression data) and $\theta_t \sim \text{normal}\left(0, \sigma_\theta^2\right)$ with $\sigma_\theta^2 \sim \text{ig}\left(.1, .1\right)$ when the generating called for a unidimensional ability parameter (for the AFD data); $\boldsymbol{\tau}_h\left(j\right) \sim_{ind} \text{normal}_2\left(\boldsymbol{\mu}_j, .5\mathbf{I}_2\right)$, where $\boldsymbol{\mu}_j \sim \text{normal}_2\left(\mathbf{0}, 25\mathbf{I}_2\right)$; and $\upsilon_h\left(\mathbf{x}\right) \sim_{ind} \text{beta}\left(1, \alpha\right)$, where α was set to 2. $N_{\max}$ was set to 50.

During the analysis of each data set, I ran the MCMC sampling algorithm for 200,000 sampling iterations in order to perform Bayesian posterior estimation. I discarded the first 100,000 samples (i.e., burn-in period) and saved every fifth sample thereafter for a total of 20,000 MCMC samples that were used for posterior inferences.

Analysis of the Verbal Aggression Data Set

Diagnostics

The trace plots of the MCMC samples corresponding to the parameters in the mDP model indicate that the samples displayed good mixing of values to represent the posterior distribution (i.e., the chain explored the support of the posterior distribution) and the parameter estimates stabilized after the burn-in period. Figure 11 contains the trace plots of the MCMC saved samples of the two-dimensional ability estimates for two different persons who completed the Verbal Aggression questionnaire. Figure 12 contains the trace plots of the MCMC saved samples

of the item category step estimates for two items, each item corresponding to a different person. The 95% MCMC half-width intervals for the person abilities had range (0.008, 0.022) and for the item category steps had range (0.008, 0.033), which suggests that the posterior mean estimates of these parameters had good precision.

Predictive Performance of the Data

I compared the predictive performance of the mDP model to each of the comparison models, using the $D(m)$ criterion. The $D(m)$ for each model based on the analysis of the Verbal Aggression data are in Table XVI, with its corresponding goodness of fit and penalty terms. Recall that, for this data set, the finite-mixture PCM was not one of the comparison models because the Winmira software cannot fit a model in which the person abilities are multidimensional. The mDP model bested the PCM, the next best model, by approximately 169 $D(m)$ units. The mDP model outperforming the comparisons models is consistent with the findings in the predictive performance portion of the simulation study reported in Chapter 3.

Posterior Mean Estimates of the Mixing Distribution for the Category Steps

For the mDP model, given covariates **x**, which in this study were item indicators, the posterior mean estimates of the mixing distribution $G_{\mathbf{x}}\left(\boldsymbol{\tau}\left(\mathbf{x}\right)\right)$ reveal how the respondents used the rating categories. The posterior mean and *SD* estimates of $G_{\mathbf{x}}\left(\boldsymbol{\tau}\left(\mathbf{x}\right)\right)$ for all items, which are in Table XVII, provide an overview of all mixing distributions corresponding to the item category steps. Across the 24 items, the posterior mean estimates for the category steps had range (–1.08, 4.03). This indicates that the items were relatively hard compared to the population from which the sample was drawn because the posterior mean for each of the two-dimensional

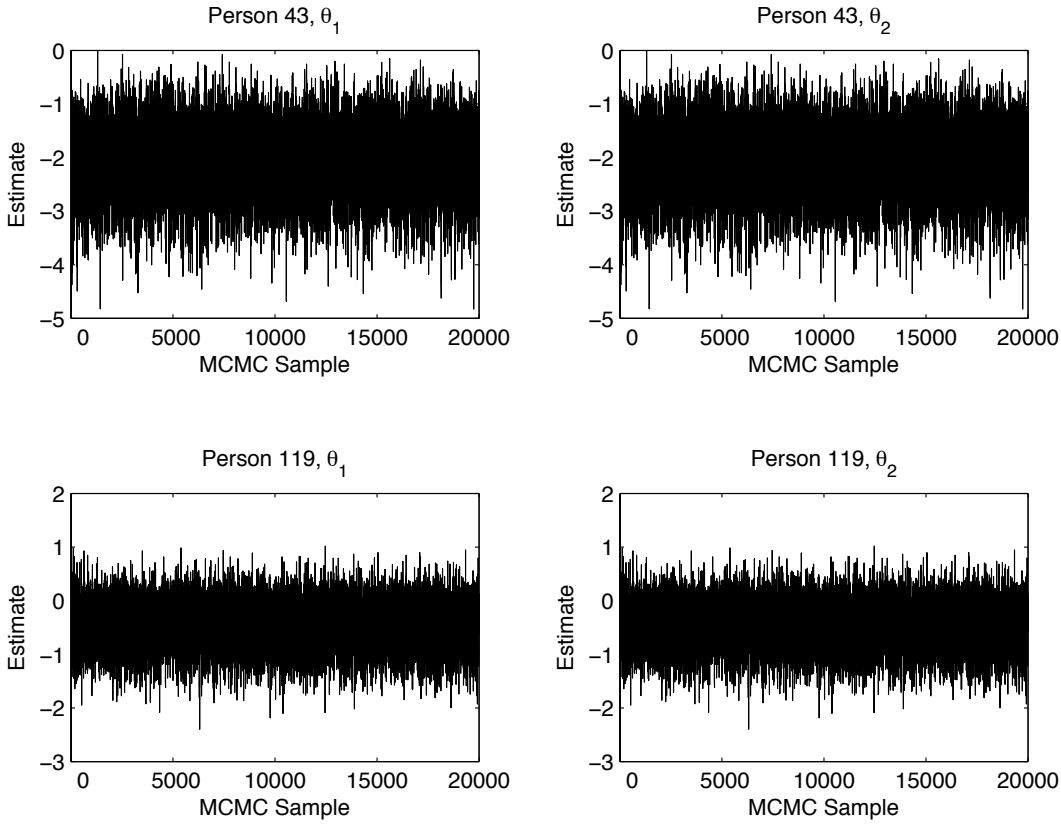

Figure 11. Trace plots of the MCMC saved samples of the two-dimensional abilities for two persons.

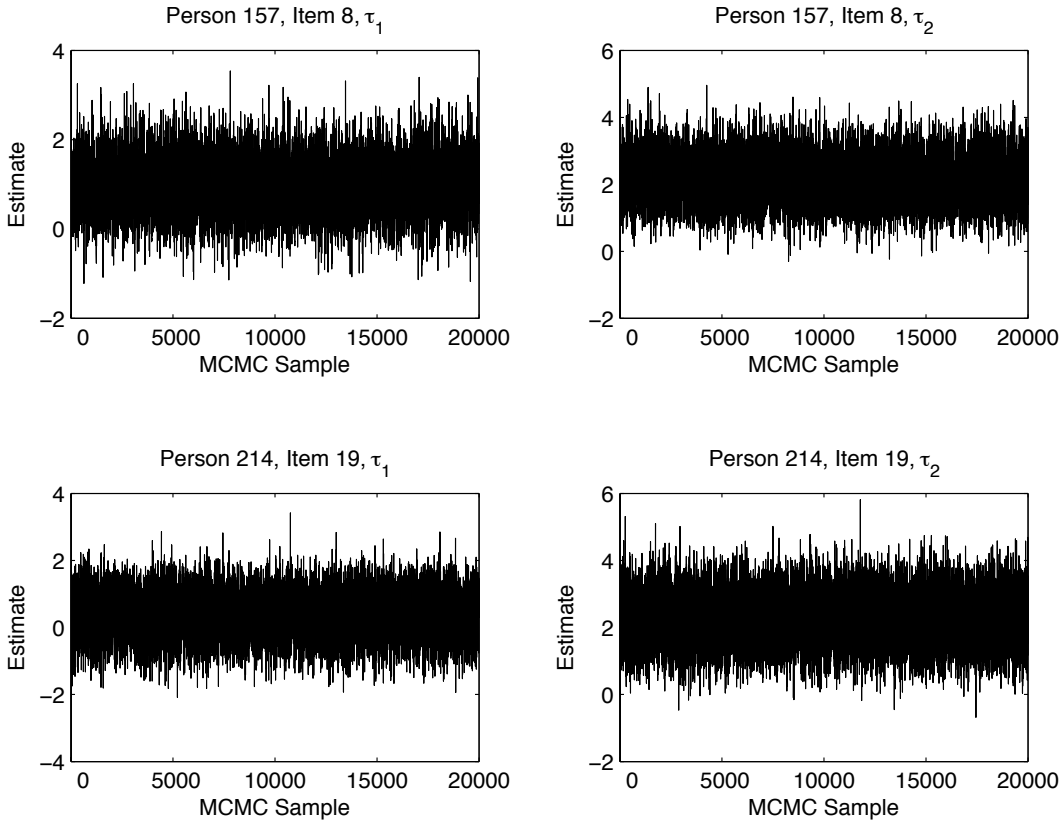

Figure 12. Trace plots of the MCMC saved samples of category steps for two items, each set corresponding to a different person.

TABLE XVI

THE *D*(*m*), GOODNESS OF FIT, AND PENALTY VALUES BY MODEL BASED ON THE ANALYSIS OF THE VERBAL AGGRESSION DATA

Model	*D*(*m*)	Goodness of Fit	Penalty
mDP	5091	2214	2877
GPCM	5260	2517	2743
GRM	5278	2527	2751
NRM	5331	2525	2806
PCM	5348	2554	2794
RSM	5361	2564	2797

Note. mDP = multiple Dirichlet process model, GPCM = generalized partial credit model, PCM = partial credit model, RSM = rating scale model, NRM = nominal response model, and GRM = graded response model. Lower value indicates better predictive performance.

person abilities was .02. That is, the respondents did not display the high level of verbal aggression associated with the higher categories. The posterior *SD* estimates had range (0.51, 0.84), with the first steps for Items 13 and 16 having estimates of 0.51 and the second step for Item 21 having an estimate of 0.84. This indicates that the posterior densities for the category steps across the items had fairly similar variances, though the second step for Item 21 had slightly larger variability in its posterior density.

A better depiction of the manner in which the respondents used the rating categories are reflected in the (marginal) posterior mean density estimates of $G_{\mathbf{x}}\left(\tau_1\left(\mathbf{x}\right)\right)$ and $G_{\mathbf{x}}\left(\boldsymbol{\tau}_2\left(\mathbf{x}\right)\right)$. The densities for Items 11, 15, and 21 of the Verbal Aggression questionnaire are in Figure 13. The densities for these items were unimodal, though the density corresponding to the second step for Item 21 had slightly flatter and wider form, reflecting the larger posterior *SD* estimate previously

TABLE XVII

POSTERIOR MEAN AND STANDARD DEVIATION (IN PARENTHESES) ESTIMATES OF THE ITEM CATEGORY STEPS PRODUCED DURING THE ANALYSIS OF THE VERBAL AGGRESSION DATA WITH THE mDP MODEL

Item	Step 1	Step 2
1	–0.50 (0.56)	–0.03 (0.55)
2	0.11 (0.55)	0.23 (0.54)
3	0.32 (0.58)	1.13 (0.68)
4	–1.08 (0.57)	0.02 (0.57)
5	–0.07 (0.52)	0.29 (0.58)
6	0.51 (0.60)	0.73 (0.59)
7	–0.16 (0.59)	1.17 (0.63)
8	0.88 (0.53)	2.13 (0.63)
9	1.64 (0.65)	2.78 (0.66)
10	–0.64 (0.63)	0.80 (0.71)
11	0.70 (0.52)	1.36 (0.63)
12	1.38 (0.66)	1.56 (0.66)
13	–0.74 (0.54)	0.40 (0.58)
14	0.09 (0.51)	0.88 (0.53)
15	1.29 (0.64)	1.73 (0.70)
16	–0.33 (0.51)	0.43 (0.60)
17	0.48 (0.52)	1.29 (0.59)
18	1.85 (0.59)	2.15 (0.67)
19	0.47 (0.66)	2.22 (0.75)
20	1.75 (0.60)	2.85 (0.65)
21	3.17 (0.64)	4.03 (0.84)
22	–0.33 (0.54)	1.10 (0.71)
23	0.73 (0.56)	1.78 (0.59)
24	2.31 (0.69)	2.75 (0.69)

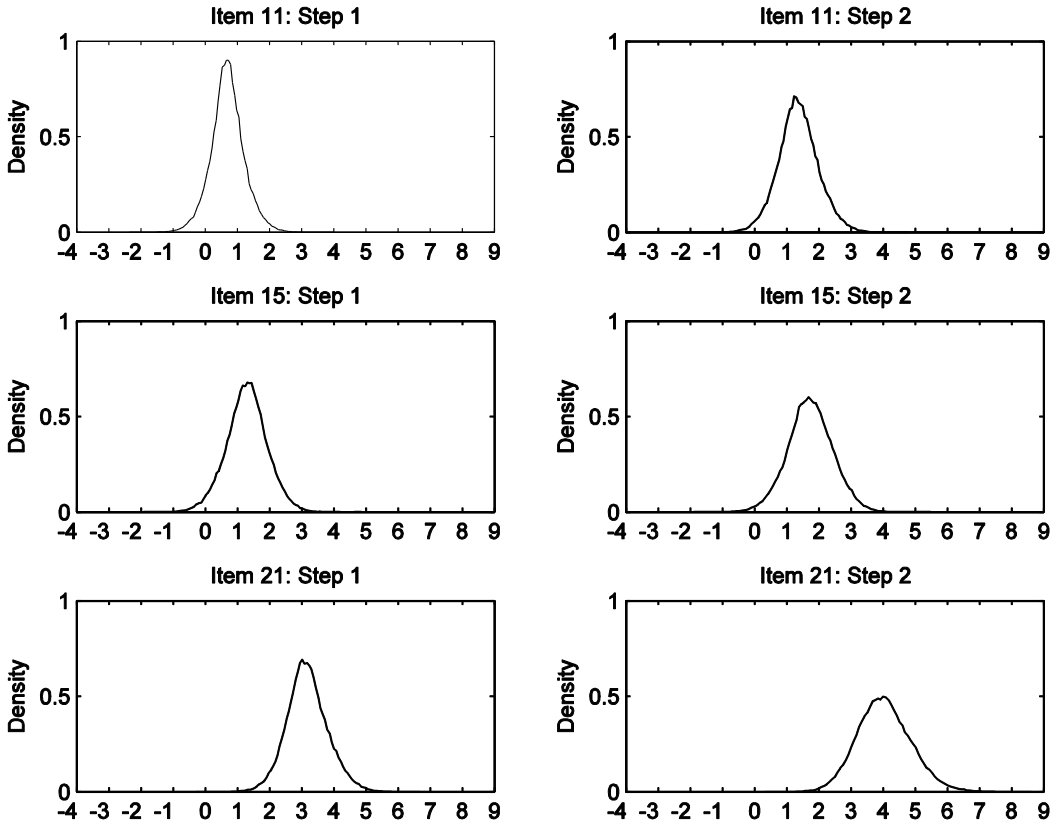

Figure 13. The marginal posterior mean density estimates of the rating category steps for three items contained in the Verbal Aggression questionnaire.

mentioned for this step. The posterior densities for the other items not presented in this figure had similar forms to Items 11 and 15, that is, unimodal and little variability.

Overview of Person Abilities

Based on the analysis of the data with the mDP model, the posterior mean estimates of the two-dimensional ability distribution were .02 for both dimensions. The posterior variance estimate of the first and second dimensions were 1.17 and 1.90, with a posterior correlation estimate of .75 between the two dimensions. This posterior correlation is similar to the estimated correlation between the two ability dimensions that the comparison models provided. The correlation estimated from the comparison models had range (.76, .79). Comparing the posterior variances of the ability distribution to the variance estimates from the comparison models is more difficult because some of the comparison models allow for the items to have different levels of discrimination (e.g., GPCM, GRM, and NRM), thus requiring those models to fix the person ability variances to 1. Therefore, I did not compare the variance estimates across the models.

Analysis of the Acculturative Family Distancing Data Set

Diagnostics

The trace plots of the MCMC saved samples of the parameters of interest, which were produced during the analysis of the AFD data set with the mDP model, showed that the parameter estimates stabilized after the burn-in period and the chains mixed well (i.e., the chain explored the support of the posterior distribution). Figure 14 contains the trace plots of the MCMC saved samples of the unidimensional ability estimates for four persons who completed the AFD questionnaire. Figure 15 contains the trace plots of the MCMC saved samples of the

item category step estimates for two items, each set of trace plots corresponding to a different person. Trace plots corresponding to the other ability and category step estimates displayed similar patterns to those displayed in these two figures. The 95% MCMC half-width intervals for the person ability estimates had range (0.008, 0.022) and for the item category step estimates had range (0.008, 0.033), which suggests that the posterior mean estimates of these parameters had good precision.

Predictive Performance of the Data

I compared the predictive performance of the mDP model to each of the comparison models, using the $D(m)$ criterion. The $D(m)$ for each model based on the analysis of the AFD data is in Table XVIII, with its goodness of fit and penalty terms. The mDP model outperformed the other models except the 3-mixture PCM. The 3-mixture PCM outperformed the mDP model by approximately 197 units. One possible reason for the better performance of the 3-mixture PCM is that three latent classes could be present in the data for all items. As previously noted, a finite-mixture IRT model assumes that the same number of latent classes apply across all items. This would be a limitation when the number of latent classes vary across items but would be the appropriate model when this assumption holds. Nevertheless, the 585 $D(m)$ unit difference between the best among the non-mixture models and the mDP model suggests that assuming that a single latent class exists across all items is inappropriate for the AFD data.

Posterior Mean Estimates of the Mixing Distribution for the Category Steps

The posterior means and *SD* estimates of $G_{\mathbf{x}}\left(\boldsymbol{\tau}\left(\mathbf{x}\right)\right)$ for all items are in Table XIX and provide a quick overview of all mixing distributions corresponding to the item category steps. Across the 24 items, the posterior mean estimates of the category steps ranged between –0.46

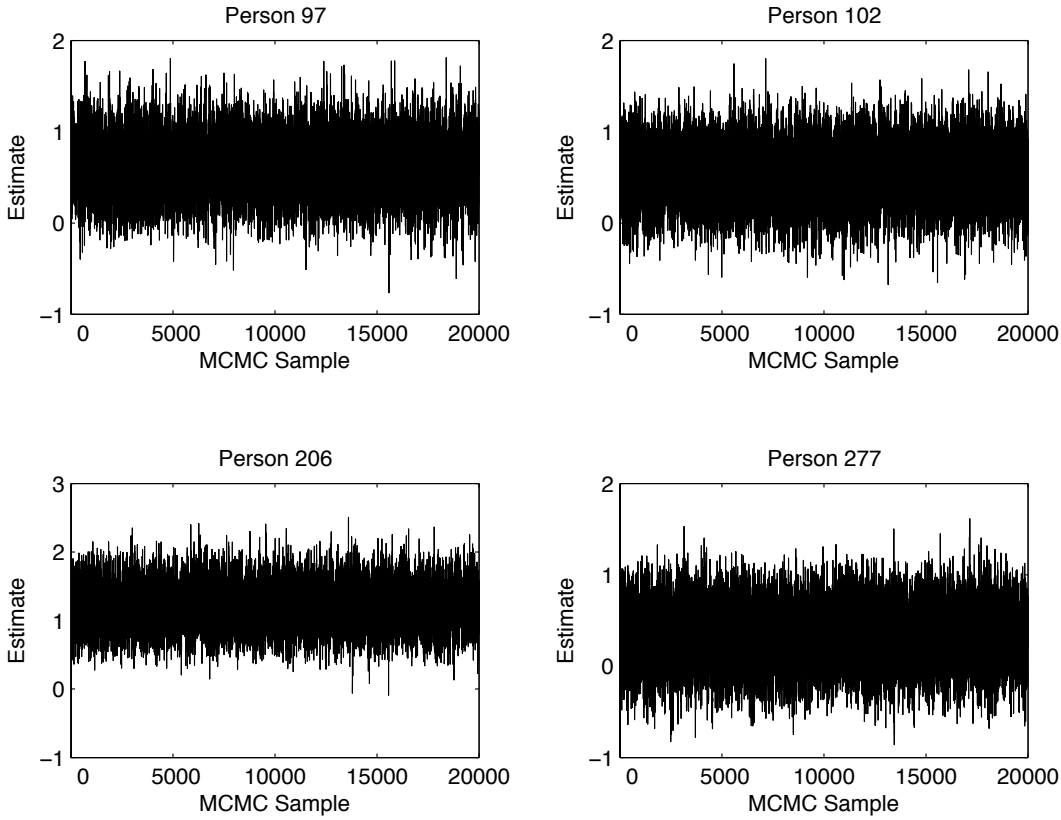

Figure 14. Trace plots of the MCMC saved samples of the unidimensional ability estimates for four persons.

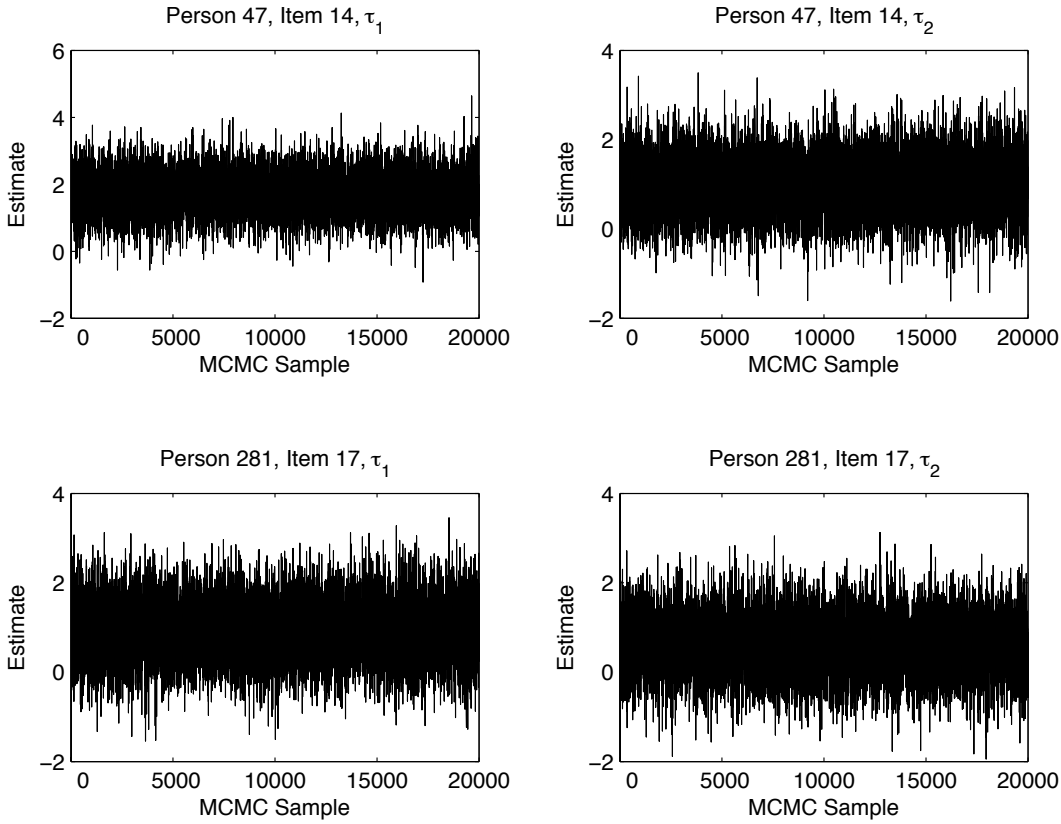

Figure 15. Trace plots of the MCMC saved samples of item category step estimates for two items, each set corresponding to a different person.

TABLE XVIII

THE $D(m)$, GOODNESS OF FIT, AND PENALTY VALUES BY MODEL BASED ON THE ANALYSIS OF THE AFD DATA

Model	$D(m)$	Goodness of Fit	Penalty
mDP	5395	2281	3114
3-mixture PCM	5198	2479	2720
GPCM	5980	2949	3031
NRM	5977	2942	3035
GRM	6007	2978	3029
PCM	6133	3016	3116
RSM	6140	3023	3118

Note. mDP = multiple Dirichlet process model, GPCM = generalized partial credit model, PCM = partial credit model, RSM = rating scale model, NRM = nominal response model, and GRM = graded response model. Lower value indicates better predictive performance. For the finite-mixture PCM, the optimal number of latent classes was identified using the AIC.

and 2.37. This indicates that the items were relatively hard compared to the population from which the sample was drawn because the posterior mean estimate of the unidimensional person ability distribution was .01. Thus, the families that completed the AFD questionnaire did not have high levels of AFD on average. The posterior *SD* estimates ranged between 0.49 and 1.43, with the first step for Item 20 having an estimate of 0.49 and the second step for Item 24 having an estimate of 1.43. This indicates that there was variability across the mixing distributions of the category steps.

The marginal posterior mean density estimates of $G_{\mathbf{x}}\left(\tau_1(\mathbf{x})\right)$ and $G_{\mathbf{x}}\left(\tau_2(\mathbf{x})\right)$ for Items 20, 23, and 24, which are in Figure 16, show the range of different forms that $G_{\mathbf{x}}\left(\boldsymbol{\tau}(\mathbf{x})\right)$ took.

TABLE XIX

POSTERIOR MEAN AND STANDARD DEVIATION (IN PARENTHESES) ESTIMATES OF THE ITEM CATEGORY STEPS PRODUCED DURING THE ANALYSIS OF THE AFD DATA WITH THE mDP MODEL

Item	Step 1	Step 2
1	2.37 (0.55)	1.65 (0.61)
2	1.85 (0.51)	1.35 (0.54)
3	2.02 (0.52)	0.69 (0.56)
4	1.51 (0.57)	0.36 (0.56)
5	1.03 (0.68)	0.55 (0.62)
6	1.17 (0.53)	0.45 (0.67)
7	0.69 (0.59)	–0.26 (0.73)
8	0.78 (0.54)	–0.46 (0.63)
9	1.85 (0.72)	0.64 (0.91)
10	0.95 (0.72)	0.53 (0.94)
11	0.89 (0.72)	0.35 (0.68)
12	1.14 (1.15)	1.64 (1.15)
13	2.03 (0.53)	1.63 (0.60)
14	1.75 (0.51)	1.12 (0.56)
15	1.76 (0.54)	1.04 (0.55)
16	1.09 (0.57)	0.35 (0.54)
17	0.92 (0.55)	0.65 (0.54)
18	0.73 (0.50)	0.83 (0.67)
19	0.98 (0.52)	0.19 (0.57)
20	0.71 (0.49)	0.32 (0.60)
21	1.97 (0.65)	0.79 (0.90)
22	0.52 (0.66)	–0.09 (0.91)
23	0.60 (0.68)	–0.01 (1.07)
24	0.52 (1.21)	1.01 (1.42)

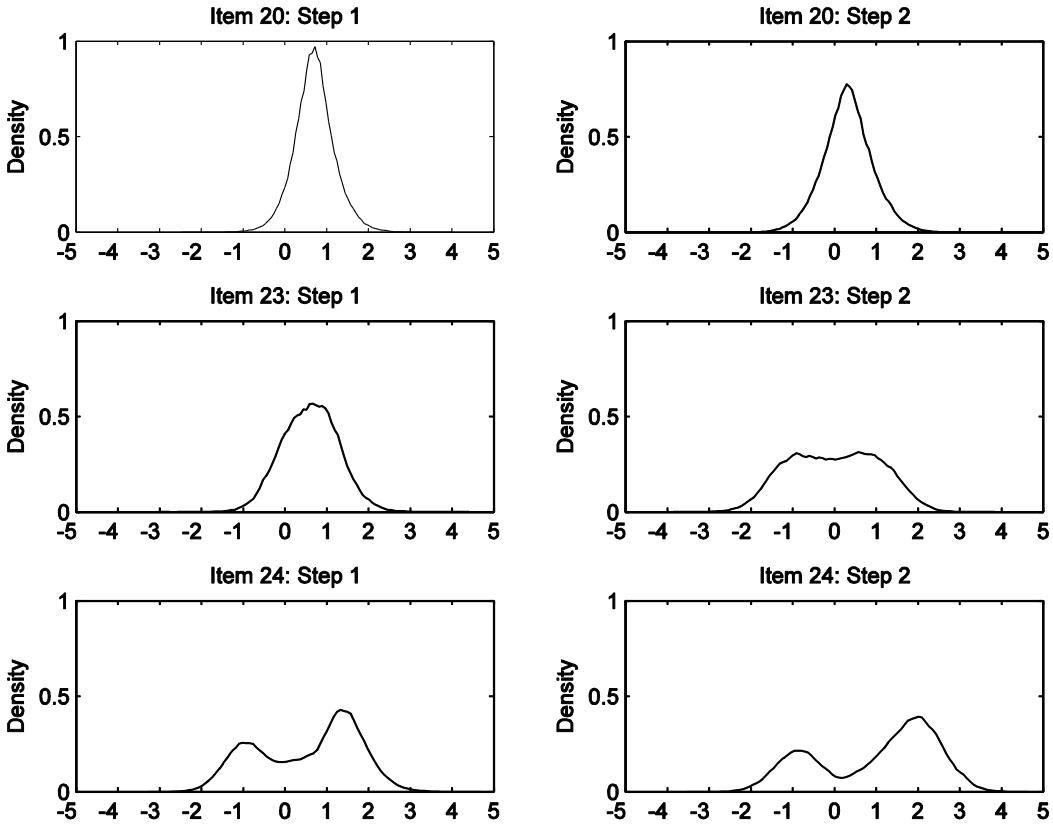

Figure 16. The marginal posterior mean density estimates of the rating category steps for three items contained in the AFD questionnaire.

Item 20's posterior densities for the category steps were unimodal with small variances. In contrast, the category step posterior densities for Item 24 were bimodal, suggesting that two latent classes were present in the data, and these classes affected the category usage across the range of the rating categories within this items (i.e., DSF resided in both steps). For Item 23, the posterior density for the first category step was unimodal while the posterior density for the second category step was slightly bimodal. This indicates that two latent classes were present in the data corresponding to the upper two categories but only a single group corresponding to the lower categories.

Person Ability Overview

Based on the analysis of the AFD data with the mDP model, the posterior mean for person ability distribution was .01, with posterior variance of 1.47. Again, straight forward comparisons of the ability distribution variance estimates across all models is difficult because some of the comparison models allow the items to have differing levels of discrimination (e.g., GPCM, GRM, and NRM), thus requiring the model to fix the variances to 1. Therefore, I did not compare the variance estimates across the models.

V. FOLLOW-UP SIMULATION ANALYSES

The marginal posterior mean estimates of the mixing distributions for the category steps in Figure 16 suggest that Item 23 had DSF between two latent classes in the categories associated with the second step, and Item 24 had DSF between two latent classes at both steps. However, the simulation study in Section III suggests that the mDP version of the DDP-RM might not be capable of detecting DIF or DSF given the conditions of the simulation study. The one instance in which a bimodal density appeared corresponded to an item that was supposed to be free of DSF, though in all other instances, the posterior densities for the random category steps were unimodal with fairly small variances. The question then arises, are the densities in Figure 16 also an anomaly, or do Items 23 and 24 indeed have DSF across two latent classes. If the latter, then another question arises. Was the 1-logit difference used for the DSF effect and the sample size conditions in the simulation study large enough for the mDP model to identify the latent classes contributing to DSF in an item?

If indeed the DSF effect and/or sample size played a role the mDP model's lack of ability to detect the DSF items in the simulation study presented in Chapter III, this issue could be analogous to power in detecting a statistically significant difference between two group means, such as via a *t*-test. To explore this possibility, I performed two follow-up analyses. The first analysis involved examining how the frequency distribution of the data corresponding to DSF items changed as a function of sample size and DSF effect. The second follow-up analysis involved generating marginal posterior mean estimates of the mixing distributions for the category steps based on data generated under a wider condition of sample sizes and DSF effects.

Frequency Distribution of the Data

For this portion of the analysis, the person abilities were assumed to be unidimensional and were drawn from a normal distribution with a mean of 0 and variance of 2.25. The sample size conditions consisted of 500; 1,000; 2,000; and 3,000. The other condition that varied was the DSF effect between two groups (with the sample size divided equally between the two groups). One item had DSF, with the following effects sizes: 1-, 2-, 2.5-, 3-, and 3.5-logit difference. Within each of the four sample sizes by six DSF conditions, I generated 1,000 data sets and averaged the frequencies. Each data set consisted of 3-point rating scale (k = 0, 1, or 2) generated for 10 items. The first nine items were free of DSF and the 10^{th} item had DSF in the second step. The generating values for the ten items are in Table XX. The value presented in the table corresponding to Item 10's second step is the overall step value for the two groups. For a given DSF effect, the effect was divided by two and then subtracted from this overall step value for Group 1's generating step value and added to this overall step value for Group 2's generating step value. For example, when the DSF effect was 2, Group 1's generating step value was –0.50 and Group 2's generating step value was 1.50.

The average frequencies across the 1,000 generated data for Item 10 (the DSF item) are in Table XXI. Because the DSF resided in the second step, the key values are in the rows corresponding to category 2. When N = 500, under the 1-logit DSF condition, the difference in average frequencies of a score of 2 between the two groups was 44. This difference in frequency in the categories corresponding to where the DSF lies could be the reason why the mDP model could not identify the problematic items through the marginal posterior predictive densities for the item category steps. When the DSF effect was 1-logit, the difference in the sheer number of persons generated with values corresponding to category 2 between the two groups progressively

TABLE XX

GENERATING CATEGORY STEP VALUES FOR THE ITEMS USED IN BOTH FOLLOW-UP ANALYSES

Item	Step 1	Step 2
1	–2.30	–1.30
2	–1.85	–0.85
3	–1.30	–0.30
4	–0.90	0.10
5	–0.50	0.50
6	–0.10	0.90
7	0.30	1.30
8	0.85	1.85
9	1.30	2.30
10	–1.50	Group 1 = 0.50 – (DSF effect/2) Group 2 = 0.50 + (DSF effect/2)

Note. The DSF resided in the second step of Item 10. The value of 0.50 was the overall starting value for the two groups. When the DSF effect was 2, Group 1's generating value was –0.50 and Group 2's generating value was 1.50.

increased as the sample size increased. For instance, when $N = 1{,}000$, the difference increased to 78. When $N = 2{,}000$, the difference increased to 148. Finally, when $N = 3{,}000$, the difference increased to 240. This trend of difference in the frequency of scores between the two groups as a function of sample size appeared under different DSF effect conditions as well.

Within each sample size condition, the difference between two groups with respect to frequency of scores of 2 also increased as the DSF effect increased. When $N = 500$, there was a difference of 115 in frequency of scores of 2 between the two groups when the DSF effect was 3 logits, and the difference increased to 130 when the DSF effect increased to 3.5 logits. This

TABLE XXI

AVERAGE FREQUENCY DISTRIBUTION (PROPORTION WITHIN GROUP IN PARENTHESES) OF GENERATED RESPONSES BY GROUP AS A FUNCTION OF SAMPLE SIZE AND DSF EFFECT

			DSF Effect					
N	Group	Category	DSF Free	1 Logit	2 Logits	2.5 Logits	3 Logits	3.5 Logits
500	1	0	50 (.20)	46 (.18)	42 (.17)	40 (.16)	38 (.15)	35 (.14)
		1	100 (.40)	84 (.34)	68 (.27)	61 (.24)	54 (.22)	47 (.19)
		2	100 (.40)	120 (.48)	140 (.56)	149 (.59)	158 (.63)	167 (.67)
	2	0	52 (.21)	55 (.22)	57 (.23)	58 (.23)	59 (.24)	60 (.24)
		1	103 (.41)	119 (.48)	134 (.54)	141 (.57)	148 (.59)	153 (.61)
		2	95 (.38)	76 (.30)	58 (.23)	50 (.20)	43 (.17)	37 (.15)
1000	1	0	110 (.22)	103 (.21)	95 (.19)	91 (.18)	86 (.17)	81 (.16)
		1	200 (.40)	168 (.34)	138 (.28)	123 (.25)	109 (.22)	96 (.19)
		2	190 (.38)	229 (.46)	267 (.53)	286 (.57)	304 (.61)	322 (.64)
	2	0	108 (.22)	114 (.23)	118 (.24)	121 (.24)	122 (.24)	123 (.25)
		1	204 (.41)	235 (.47)	265 (.53)	278 (.56)	290 (.58)	302 (.60)
		2	188 (.38)	151 (.30)	117 (.23)	102 (.20)	88 (.18)	75 (.15)
2000	1	0	215 (.22)	200 (.20)	184 (.18)	175 (.17)	164 (.16)	154 (.15)
		1	412 (.41)	348 (.35)	285 (.28)	254 (.25)	227 (.23)	200 (.20)
		2	373 (.37)	452 (.45)	531 (.53)	571 (.57)	609 (.61)	646 (.65)
	2	0	216 (.22)	226 (.23)	235 (.23)	240 (.24)	242 (.24)	245 (.24)
		1	406 (.41)	470 (.47)	530 (.53)	556 (.56)	582 (.58)	606 (.61)
		2	379 (.38)	304 (.30)	235 (.24)	204 (.20)	176 (.18)	150 (.15)
3000	1	0	329 (.22)	308 (.21)	282 (.19)	268 (.18)	254 (.17)	238 (.16)
		1	615 (.41)	519 (.35)	424 (.28)	380 (.25)	337 (.22)	297 (.20)
		2	556 (.37)	674 (.45)	794 (.53)	852 (.57)	909 (.61)	965 (.64)
	2	0	332 (.22)	350 (.23)	363 (.24)	369 (.25)	373 (.25)	378 (.25)
		1	622 (.41)	716 (.48)	801 (.53)	839 (.56)	877 (.58)	908 (.61)
		2	546 (.36)	434 (.29)	335 (.22)	292 (.19)	250 (.17)	214 (.14)

Note. One item was treated as having DSF. Values are averaged across 1,000 data sets. The DSF effect was in the step separating categories 1 and 2.

pattern appeared with other sample sizes as well. These frequencies indicate the role sample size and DSF effect size play in the number of persons being impacted by DSF.

Posterior Mean Estimates of the Mixing Distribution for the Category Steps

Within each of the sample size (500; 1,000; 2,000; and 3,000) by DSF effect (2-, 2.5-, 3-, and 3.5-logit difference) conditions, I analyzed a simulated data set generated for the previous section with the mDP model. I assigned the same proper priors on the estimated parameters of the mDP model that were assigned in the simulation study in Section III and the analysis of real-life data sets in Section IV. That is, $\theta_t \sim \text{normal}\left(0,\sigma_\theta{}^2\right)$ with $\sigma_\theta{}^2 \sim \text{ig}\left(.1,.1\right)$; $\boldsymbol{\tau}_h\left(j\right) \sim_{ind} \text{normal}_2\left(\boldsymbol{\mu}_j,.5\mathbf{I}_2\right)$, where $\boldsymbol{\mu}_j \sim \text{normal}_2\left(\mathbf{0},25\mathbf{I}_2\right)$; and $\upsilon_h\left(\mathbf{x}\right) \sim_{ind} \text{beta}\left(1,\alpha\right)$, where α was set to 2. $N_{\max}$ was set to 50. I ran the MCMC sampling algorithm for 110,000 sampling iterations in order to perform Bayesian posterior estimation. I discarded the first 50,000 samples (i.e., burn-in period) and saved every third sample thereafter for a total of 20,000 MCMC samples that were used for posterior inferences.

I examined the (marginal) posterior mean density estimates of $G_{\mathbf{x}}\left(\tau_1\left(\mathbf{x}\right)\right)$ and $G_{\mathbf{x}}\left(\boldsymbol{\tau}_2\left(\mathbf{x}\right)\right)$ to determine whether the mDP model was capable of detecting DSF between two latent classes as the sample size increased, holding the DSF effect constant, and as the DSF effect increased, holding the sample size constant. All items were treated as random for this portion of the study. Figure 17, 18, 19, 20, and 21 contain the marginal posterior densities of the category steps for Item 10 (the DSF item) for each sample size when the DSF effect between the two groups in the second step was 1, 2, 2.5, 3, and 3.5 logits, respectively.

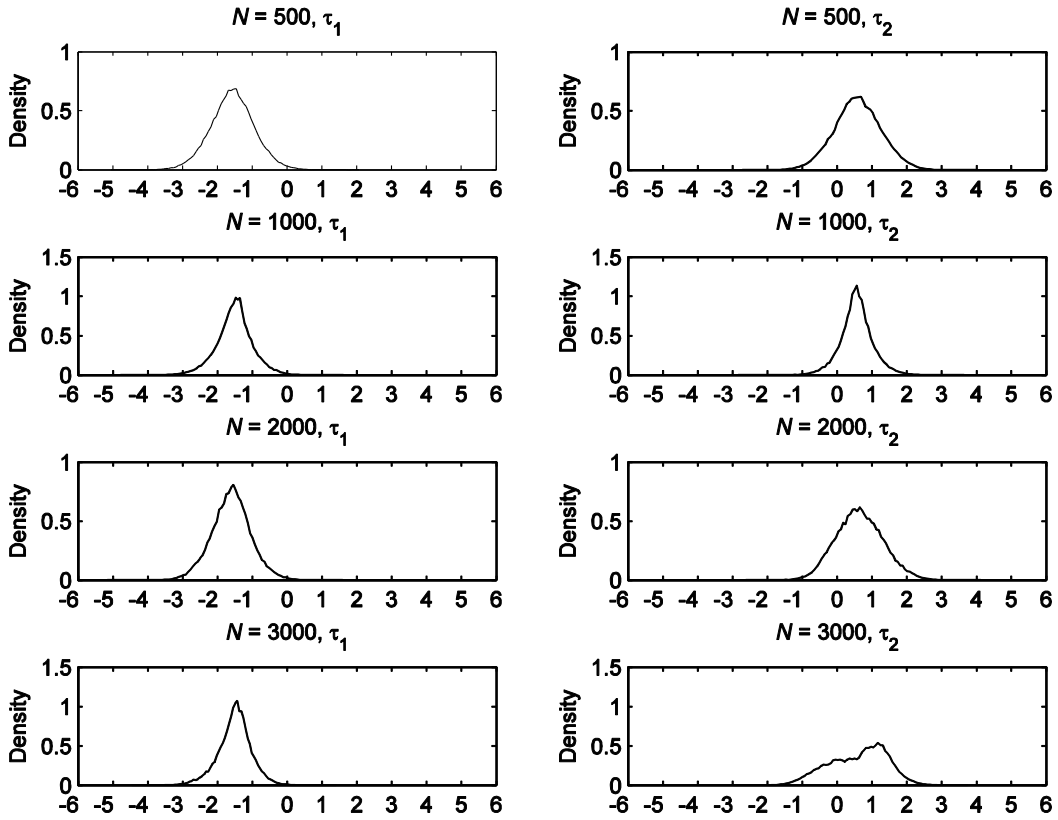

Figure 17. The marginal posterior mean density estimates of the rating category steps for Item 10 (the DSF item) by sample size for the data sets corresponding to a 1-logit DSF effect in the second step.

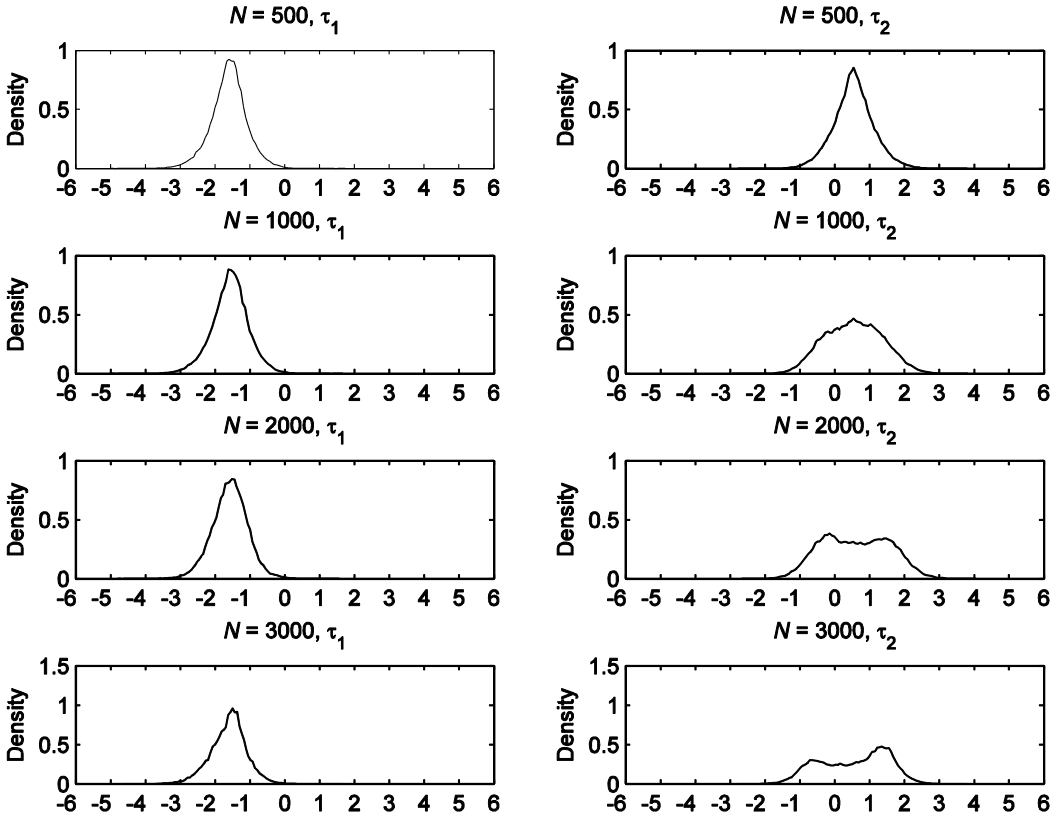

Figure 18. The marginal posterior mean density estimates of the rating category steps for Item 10 (the DSF item) by sample size for the data sets corresponding to a 2-logit DSF effect in the second step.

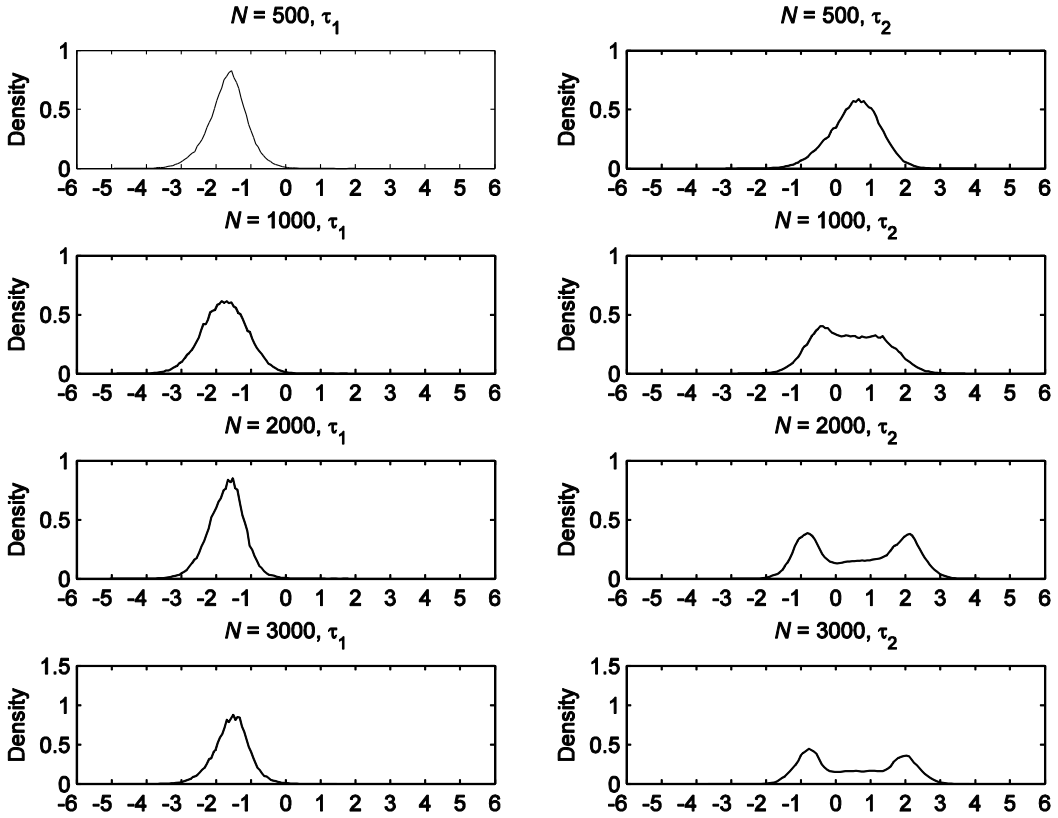

Figure 19. The marginal posterior mean density estimates of the rating category steps for Item 10 (the DSF item) by sample size for the data sets corresponding to a 2.5-logit DSF effect in the second step.

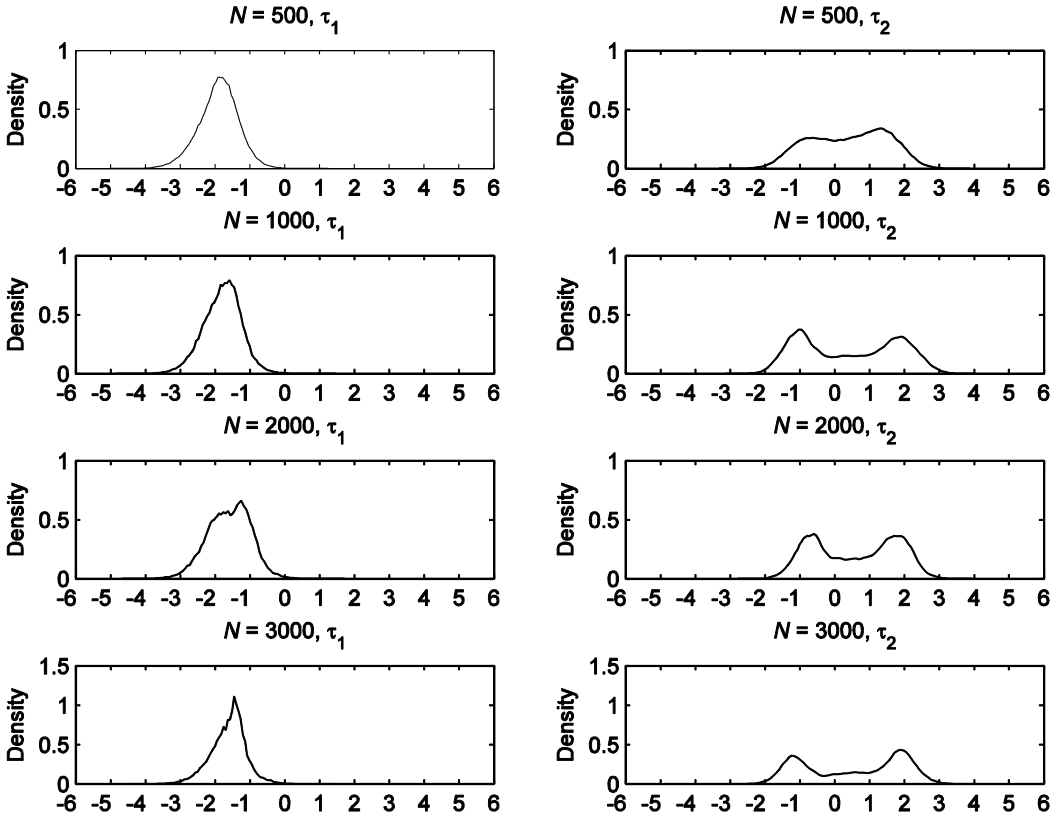

Figure 20. The marginal posterior mean density estimates of the rating category steps for Item 10 (the DSF item) by sample size for the data sets corresponding to a 3-logit DSF effect in the second step.

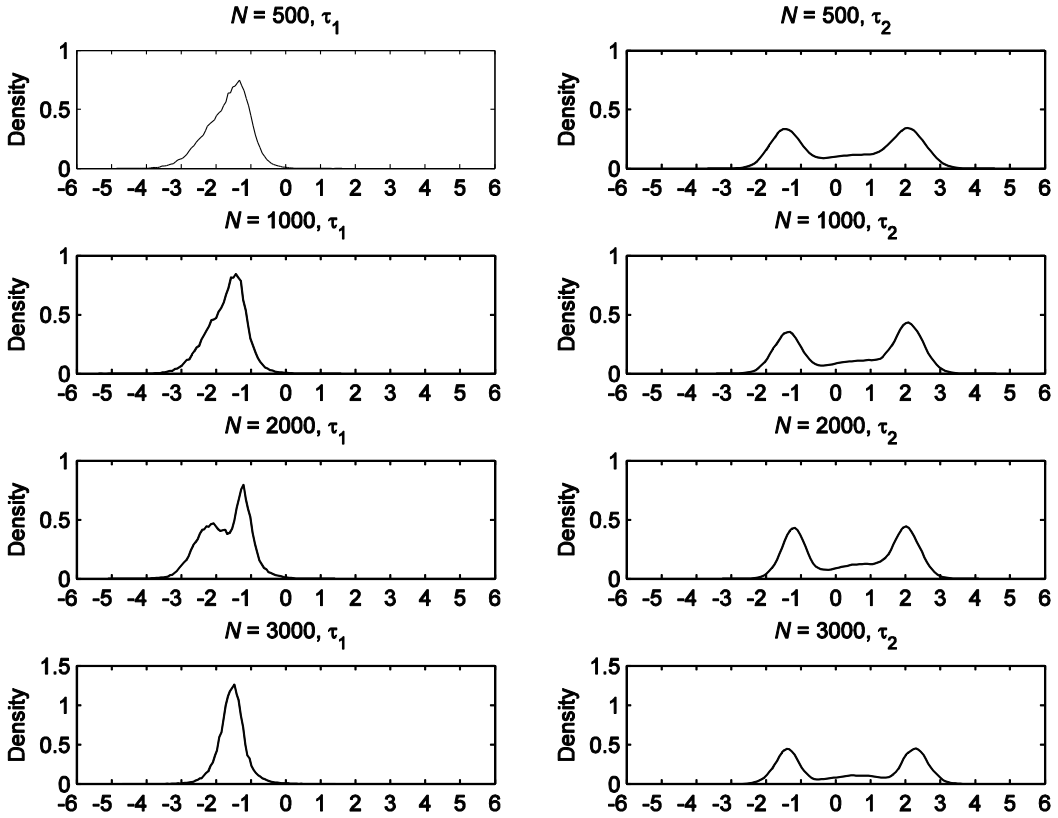

Figure 21. The marginal posterior mean density estimates of the rating category steps for Item 10 (the DSF item) by sample size for the data sets corresponding to a 3.5-logit DSF effect between two groups in the second step.

The posterior densities across these five figures indeed suggest that, as sample size and the DSF effect increases, the effectiveness of the posterior densities to capture the correct number of latent classes contributing to DSF increases. When the DSF effect was 1 logit, the posterior density of the category step infected with DSF only became nonnormal when the sample size reached 3,000, and even then, the density was not truly bimodal. When the DSF effect was 2 logits (Figure 18), at $N = 500$, the posterior density for the second step was still unimodal, thus indicating that DSF was not present in this step. As the sample size increased, however, the variability in the posterior densities corresponding to the second step increased, with the density corresponding to $N = 2{,}000$ showing slight bimodality and $N = 3{,}000$ displaying a slightly more defined bimodal form. The densities in this figure suggest that, for a DSF effect of 2 logits between two groups, a sample size of 500 is not large enough for the mDP model to detect the DSF. However, as the sample size increases, the marginal posterior predictive densities of the mixing distributions that the mDP model produces becomes more effective at capturing the DSF, with the number of modes corresponding to the number of latent classes contributing to the DSF. Within this DSF condition, the posterior densities corresponding to the first step were unimodal, which should be the case because the first step was free of DSF. This general pattern held for each DSF condition, as indicated by the five figures. The exception occurred in the posterior density corresponding to the first step for $N = 2{,}000$ and DSF condition of 3.5 logits (Figure 21). The density for this step had a slight bimodal form, but one mode was definitely larger than the other, thus suggesting that one value could still be appropriate to represent the data corresponding to the step. This density's bimodal form did not resemble the bimodal form depicted in the density corresponding to the second step. This shape could also represent the random data generating process for this data set.

The densities in Figure 17, 18, 19, 20, and 21 indicate the role DSF magnitude plays in the mDP model's ability to identify DSF items, controlling for sample size. The densities corresponding to the sample size condition of $N = 500$ across these five figures reveal that, as the DSF effect increased, the variability in the density corresponding to the second step also increased. In fact, for $N = 500$, when the DSF effect reached 3 logits, the density for the second step was slightly bimodal (Figure 20), and at an effect of 3.5 logits, the density for the second step was clearly bimodal (Figure 21). The general pattern of the form of the density becoming bimodal as the DSF effect increased occurred in the other sample size conditions as well.

The follow-up analyses show that the marginal posterior predictive densities of the mixing distributions produced by the mDP model can identify when DSF occurs between two latent classes, given that a certain DSF effect and/or sample size is reached. The follow-up analyses also indicate that the difference in the actual frequency of persons corresponding to a category score associated with the DSF step between the two groups (compared to difference in proportional distribution between the two groups) is more important. This is supported by the finding that, as the sample size increased, the DSF effect appeared in the posterior densities. Recall that, for the DSF item, the proportion of cases receiving a score associated with the step where the DSF resided remained relatively the same across the sample size condition (see Table XXI). Yet, the increase in sample size, which corresponded to an actual difference in average frequency of cases receiving a score associated with the step where the DSF existed, led to an improvement in the density reflecting the DSF condition.

VI. DISCUSSION

Summary

The work performed in this thesis was undertaken to address issues related to examining the technical properties of rating scale items for invariance across subgroups of persons. In Chapter 1, I distinguished between DIF and DSF, two forms of item property noninvariance. An item is said to have DIF when group membership (either known or latent) equally affects all category steps for an item. DSF, on the other hand, occurs when group membership has a differential effect on the category steps within an item. This differential effect is what makes examining items for DSF more challenging than examining items for DIF. To detect DSF, each category step must be examined within an item.

I then reviewed the common approaches within IRT used to examine items for DIF and DSF. The traditional approach within IRT is to include group membership information (e.g., gender or race) into the model to examine the effect of such variable(s) on the category steps. Unfortunately, this approach requires some decision on the manner in which a common metric across all groups examined is established (e.g., anchoring on a subset of items), and the chosen method could have ramifications on the effectiveness in detecting DIF/DSF (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988; Wang, 2004; 2008). This approach also ignores the possibility that latent groups, or latent classes, are present in the data that could also be contributing to DIF/DSF.

Finite-mixture IRT models were developed to address the issues of detecting when DIF/DSF occurs across latent classes. Unfortunately, these models require the number of mixture components to be specified to a finite value, the same number of mixture components are used to describe the data for all items, and this number is assumed to apply across all ability dimensions

when the latent trait space is multidimensional. Such assumption can easily be violated, such as when a subset of items is free of DIF/DSF while another subset has DIF/DSF across three racial groups. In the data corresponding to the first subset of items, only one mixture component needs to be specified while in the data corresponding to the second subset of items, three mixture components are required. This assumption can also be inappropriate when the set of items that measures one dimension of the latent trait space is free of DIF/DSF while the set of items that measures another dimension of the latent trait space has DIF/DSF.

These limitations with the multiple-group and finite-mixture IRT models in performing DSF analysis is what inspired the introduction of a new type of Bayesian nonparametric IRT model within the DDP-RM framework. In the model presented in this study, the covariate-dependent mixing distribution is formed via an mDP. The mDP model does not require group membership to be known. That is, it seeks to identify latent classes that could be contributing to DIF/DSF in an item. It does so without assuming that the same set of mixture components form the mixing distribution of the category steps across all items, which addresses the limitation of the finite-mixture IRT models. Another benefit of forming the mixing distribution via the mDP, and DDP in general, is that the mixing distribution can take on different forms, such as unimodal with small variance and bimodal with very large variance (which would be the case when an item has DSF between two groups), and the mixing distribution can vary across a set of covariates. This approach does not make the assumption that the true form of the distribution of each the category step is normally distributed, which many parametric random effects models make. Another feature of the mDP model is that it does not require a common set of items to serve as anchor items to establish a common metric across possible known or latent groups, as often done with the multiple-group and finite-mixture IRT models. The hyper priors for the

mixing distribution establish a link across all persons, which in turn leads to a link across all latent classes. Because a subset of items do not have to serve as anchors, the mDP model can simultaneously examine all items for DSF rather than only those items that do not serve as anchors.

The simulation study in Chapter III, the analysis of real-life data sets in Chapter IV, and the follow-up analyses in Chapter V display the conditions under which the mDP model can identify when DSF in a category step occurs across two latent classes. As the simulation study in Chapter III showed, the mDP model outperformed the other traditional IRT models in terms of predictive performance of the data. In this portion of the simulation study, data sets were generated to reflect conditions in which none of the items had DSF while other data sets were generated to have DSF for a subset of items. After accounting for model complexity, the mDP model was a better fit for the data than the other used to analyze the data in this simulation study, which included a finite-mixture IRT model. Unfortunately, given the simulation conditions, the posterior predictive densities for the category steps did not flag the items that had DSF, as others have found (Fujimoto & Karabatsos, 2014). There are a few possible reasons for this discrepancy in findings. First, in the other study, the DSF item had a 2-logit difference between two groups while in the simulation study performed in Chapter III, the effect was 1-logit difference. Second, the other study subjected only two items to the covariate-dependent infinite mixtures while in the present study, all items were treated as random. Third, the other study generated data for 3,000 cases while the largest number of cases in this portion of the study was 800. Finally, the mixing distributions were formed through different processes: the other study relied on a modified version of the local DP while this study utilized the mDP.

The mDP model's lack of ability to detect the DSF item within the given conditions in Chapter III and greater predictive performance of the simulated data in relation to the comparison models suggest that the gains in predictive performance was most likely because the mDP model accounted for the random variation that occurs in the response process by treating the category steps as random.

Based on the analysis of the real-life data sets, discrepancy in terms of predictive performance between the mDP model and the finite-mixture PCM appeared. The results of the AFD data analysis revealed that the 3-mixture PCM bested the mDP model with respect to predictive performance after accounting for model complexity. As previously noted, the finite-mixture IRT models assume that the specified number of latent classes in the data applies across all items, which is a limitation when the number of latent classes actually varies across items. This assumption is actually appropriate when the same number of specified latent classes is present in the data for all items. It could be the case, then, that three latent classes indeed are present in the AFD data for all items, which would make the 3-mixture PCM the most appropriate. Although the mDP model accounts for the random process in the response process, which could be why the goodness of fit (GF) is better for the mDP (GF = 2,281) than the 3-mixture PCM (GF = 2,470) (please see Table XVIII), the mDP model incorporates many more parameters to accomplish this goodness of fit than the finite-mixture PCM, which is reflected in the penalty term. The mDP model and the 3-mixture PCM had penalty terms of 3,114 and 2,720 units, respectively. These results possibly suggest that, if the assumptions of the finite-mixture IRT models are met, then the added complexity of the mDP model does not lead to greater predictive gains.

Another finding that resulted from the analysis of the AFD data is that the posterior predictive densities of the category steps for Items 23 and 24 followed a bimodal form rather than a unidimensional distribution with small variance. These bimodal densities were similar to the density for one item's steps in the simulation study in Chapter III, though that item was a DSF item. The results of the simulation study in Chapter III and the analysis of the real-life data sets in Chapter IV led to the question, are the bimodal densities that the mDP model produced during the analysis of the AFD data because the items had a large enough DSF for the densities to capture or was a false positive observed with the one density in Chapter III?

As previously noted, others showed that the marginal posterior mean predictive density of the mixing distribution was capable of revealing DSF when the magnitude of the DSF was 2 logits. Thus, it is possible that the magnitude of the DSF for these items in the AFD data was sufficient for the posterior densities to capture the DSF effect, while the magnitude of 1-logit difference between two groups used in the simulation study in Chapter III was insufficient. Thus, to gain a better understanding of the role the magnitude of the DSF and sample size plays in the mDP model's ability to detect when DSF occurs between two latent classes, I performed follow-up simulation studies for which the results were presented in Chapter V.

The follow-up analyses in Chapter V highlighted the two variables that play a role in whether the posterior predictive densities of the category steps can correctly reflect the DSF condition. Holding sample size constant, as the magnitude of the DSF increased, the variability in the densities increased and the form became bimodal; and holding the DSF effect constant, as the sample size increased, the variability in the densities also increased and the form also became bimodal. These findings indicate that the posterior predictive densities produced with the mDP model can indeed detect DSF in an item, given that a certain magnitude of DSF or sample size is

met. It accomplishes this without having to specify the number of latent classes and not restricting the same mixture components to describe the data across all items. The finding of the role of sample size and magnitude of DSF is analogous to the issue of power in detecting a statistically significant difference between two group means, such as through a *t*-test. In other words, when the mean difference between two groups is constant, a larger sample size increases the chances of obtaining a statistically significant difference. Likewise, for a given sample size, as the difference in two group means increases, so does the probability of detecting a statistically significant difference between the two groups.

Another finding from the follow-up analyses is that the sheer number of persons affected by the DSF appears to play a greater role than the proportion of persons within group that are affected by the DSF, as the average of the frequencies of scores for the DSF item suggest. That is, between the two groups, the sheer difference in the number of persons receiving scores associated with the categories where the DSF resides after controlling for ability level(s) appears to be a factor. This echoes the previous findings of the role sample size plays as a function of DSF effect. As the numbers corresponding to the categories differ between the two groups, more information is provided to estimate the posterior densities, which in turn is a likely reason why the densities corresponding to the DSF items move away from normality as the sample size increases, holding the DSF effect constant and the effect is not 0.

Limitations of the Study

Although the mDP model showed that it could identify the items with DSF, some limitations of the study should be noted. The generating item category step values used in the simulation portions of this study (Chapter III) were targeted to the ability distribution rather than randomly generating the values in order to study the effectiveness of the mDP model in detecting

the problematic items. In the follow-up analyses (Chapter V), the generating values for the category steps were also matched to the ability distribution. That is, the average value for the second step, which was where the DSF resided, was near the mean of the ability where most of the cases belonged. Also, this study only examined whether the mDP model could detect DSF when the rating scale contained three categories. The simulation conditions in this study consisted of DSF between two latent classes. In real applications, it is likely that some items could have DSF across three or more latent classes. Finally, this study did not explore whether latent classes exist within known groups. In real world applications, it is often of interest to examine whether gender contributes to DSF in an item, and it could be of interest to determine whether latent classes within each gender contribute further to DSF in an item. The mDP model presented here is general enough to accommodate such known group characteristics, but this aspect was not explored.

Future Directions and Modeling Extensions

In the immediate future, the issues noted as limitations could be addressed. That is, the item parameters can be randomly generated and then the effectiveness of detecting the DSF items through the posterior predictive densities of the category steps produced with the mDP model could be investigated. Additionally, the effectiveness of the mDP model in detecting problematic items could be examined when the category step where the DSF resides is off target from the mean of the person ability distribution. For instance, the starting point for the generating category step (i.e., the average of the generating category step value for the two groups) where the DSF resides could be two standard deviations above the mean of the ability distribution. In this case, given the findings in the follow-up simulation portion of this study (Chapter V), an even larger sample size should be required before the posterior densities can

capture the DSF compared to when the average generating step value is near the mean of the ability distribution. That is because, as the category step value moves away from the mean of the ability distribution, fewer cases within each group will be assigned to the categories separated by this step. Recall that the follow-up analyses suggest that the sheer difference in the number of cases assigned to the categories corresponding to the DSF plays a factor. Given that fewer persons will receive scores associated with a category step that is further away from the mean of the ability distribution, a larger sample size should be required before the difference in the number of cases receiving the scores corresponding the step where the DSF resides reaches a level where the posterior density could reveal the DSF.

In addition to addressing these limitations, another possible investigation could include examining whether the mDP model is more effective at recovering the generating ability values compared to other traditional IRT model when the DSF effect is strong enough so that the posterior predictive densities indicate the presence of DSF. For this thesis, the recovery of the generating ability values was only explored under conditions when the DSF could not be detected through the posterior densities (Chapter III). The recovery of the generating values was not explored in the follow-up analyses (Chapter V), where the DSF could be detected. It is possible that, when the densities can reveal the DSF effects, gains in recovery of the generating ability values could be achieved with the mDP model compared to the traditional IRT models included in this study.

With respect to modeling extensions to the mDP model, one possibility is to allow the true form of the ability distribution to be modeled nonparametrically. In this study, a normal or multivariate normal distribution was chosen as the prior for the true distributional form of the abilities. Within the mDP specification of the DDP-RM, another DP could be assigned on the

distribution of abilities $\left(\text{i.e., } \theta_t \sim DP\left(\alpha, G_{0_\theta}\right), \text{ with } G_{0_\theta} = \text{normal}\left(0, \sigma_\theta^2\right)\right)$ or a DDP when the abilities are multidimensional $\left(\text{i.e., } \boldsymbol{\theta}_t \sim DDP\left(\alpha_{\mathbf{x}}, G_{\mathbf{x}0_\theta}\right), \text{ with } G_{\mathbf{x}0_\theta} = \text{normal}_Q\left(\mathbf{0}, \Sigma_\theta\right)\right)$, where Q is the number of dimensions in the latent trait space. Within the mDP model, assigning a DP prior on the distribution of abilities would keep the mixing distribution independent from the mixing distributions of the category steps. That is, the ability side and the item side will remain independent from each other, which would be consistent with IRT models that assume an additive effect between the ability and item parameters. The utility of modeling the abilities nonparametrically is that, in most real life applications, the means of the abilities for the latent classes should differ, and this difference should be taken into consideration. Failing to do so could have grave implications for detecting DIF/DSF items (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988; Wang, 2004; 2008). Modeling the ability distribution nonparametrically could account for variation in the ability means across the latent classes while examining the item category steps for DSF.

Another possible future study could be to form more flexible weights for the infinite mixtures, which would lead to a covariate-dependent infinite-mixture IRT model that falls outside of the DDP framework. The reason to pursue this avenue is that, as the simulation study in Chapter V indicates, for smaller sample sizes, a very large DSF effect is required (e.g., about 2 times the standard deviation of the ability distribution for a sample size of 500). More flexible mixture weights could lead to a model that detects DSF across latent classes when the effect is weaker (e.g., 1 times the standard deviation of the ability distribution) and the sample size is smaller. The mixture weights in this study follow the stick-breaking weights proposed by Sethuraman (1994), and with these stick-breaking weights, a mixing distribution that follows a

DP can be formed. These stick-breaking weights could be restrictive in that they place higher support on the earlier mixture components compared to later components. This emphasis on the earlier components could be limiting a model in detecting DSF items when the effect and sample sizes are small. A mixture weight that relaxes this downward support on the latter mixture components could provide advantages, as Karabatsos and Walker (2012a) have shown with their mixture weights that are based on an infinite-ordered probits regression model with covariate dependence in the mean and variance.

A different type of mixture weight could possibly place support on nonconsecutive mixture components or support the range of the mixture components equally rather than following a progressively decreasing support on later mixture components as the stick-breaking weights do. This type of mixture weight could possibly be formed within the normalized random measures framework (James, Lijoi, & Prünster, 2009; Lijoi, Mena, & Prünster, 2005, 2007; Regazzini, Lijoi, & Prünster, 2003).

Conclusions

Examining whether the technical properties of educational and psychological test items are invariant across subgroups of persons (known or latent) is good practice. When the properties of an item are noninvariant across subgroups, it suggests that the characteristic(s) that describe the subgroups (e.g., gender and race) are part of the response process for that item. The implication of having a test comprised of noninvariant items is that the same test score does not represent the same level of the construct of interest across the groups, thus preventing direct comparison of individuals from different subgroups unless statistical adjustments are made to the test scores first. However, in order to make statistical adjustments to the scores, the problematic items must be identified.

Identifying noninvariant rating scale items (i.e., items that are polytomously scored) comes with an additional challenge that is not faced when the items are dichotomously scored. That is, with rating scale items, the person characteristic does not have to affect all scores equally. It could affect only a subset of scores, it could have a differential effect on all scores or it could affect all scores equally. Thus, with rating scale items, each category step should be examined by performing a differential step functioning (DSF) analysis. The mDP model, which is a Bayesian nonparametric IRT model introduced in this thesis, shows promise as a tool to perform such analysis. When certain conditions are met, the model is effective at identifying problematic rating scale items with category steps that have DSF across latent classes. It accomplishes this without having to specify the number of latent classes, allows the mixture components to vary across items, and all items can be treated as random, thus testing all items simultaneously. The potential of this model's approach to identifying DSF appears promising.

CITED LITERATURE

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. doi:10.1007/BF02293814

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47*(1), 105-113. doi:10.1007/BF02293856

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Mahwah, NJ: Lawrence Erlbaum.

Ankenmann, R. D., Witt, E., & Dunbar, S. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*(4), 277–300. doi:10.1111/j.1745-3984.1999.tb00558.x

Berk, R. (1982). *Handbook of methods for detecting test bias*. Baltimore, MD: Johns Hopkins University Press.

Basu, S. (2007, August). *Double Dirichlet process mixtures.* Paper presented at the Bayesian Nonparametric Regression: Theory, Methods and Applications, Issac Newton Institute for Mathematical Sciences, Camridge, United Kingdom.

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333–353. doi:10.1111/j.1745-3984.1996.tb00496.x

Chu, K.-L., & Kamata, A. (2007). Test equating in the presence of DIF items. In E. V. Smith, Jr., & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 435–451). Maple Grove, MN: JAM.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335–350. doi:10.1177/014662169301700402

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533–559. doi:10.1007/s11336-008-9092-x

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

De Jong, M. G., & Steenkamp, J. B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika, 75*(1), 3–32. doi:10.1007/S11336-009-9134-Z

De Jong, M. G., Steenkamp, J. B. E. M., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*(2), 260–278. doi: 10.1086/518532

Diebolt, J., & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 56*(2), 363-375. Retrieved from http://www.jstor.org/stable/2345907

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A programmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive assessment: Issues in constructed response, performance testing, and portfolio assessment* (pp. 135–165). Mahwah, NJ: Lawrence Erlbaum.

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

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 209-230. Retrieved from http://www.jstor.org/stable/2958008

Flegal, J. M., & Jones, G. L. (2011). Implementing Markov chain Monte Carlo: Estimating with confidence. In S. P. Brooks, A. E. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 175-197). New York, NY: Chapman & Hall/CRC.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*(4), 309–326. doi:10.1177/01466219922031437

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315–332. doi: 10.1111/j.1745-3984.1996.tb00495.x

Fujimoto, K. A., & Karabatsos, G. (2014). Dependent Dirichlet process rating model. *Applied*

Psychological Measurement, *38*(3), 217–228. doi:10.1177/0146621613512018

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*(365), 153–160. doi:10.2307/2286745

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological), 56*(3), 501–514. Retrieved from http://www.jstor.org/stable/2346123

Gelfand, A. E., & Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, *85*(1), 1–11. doi:10.1093/biomet/85.1.1

Gelfand, A. E., & Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association, 94*(445), 247–253. Retrieved from http://www.jstor.org/stable/2669699

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.

Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. New York, NY: Springer.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer-Nijhoff.

Hauck, W. W. (1979). The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics, 35*(4), 817–819. doi:10.2307/2530114

Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica, 52*(2), 271–320. Retrieved from http://www.jstor.org/stable/1911491

Hedeker, D., Berbaum, M., & Mermelstein, R. (2006). Location-scale models for multilevel ordinal data: Between- and within-subjects variance modeling. *Journal of Probability and Statistical Science, 4*(1), 1–20. Retrieved from http://www.i-tel.com.tw/jpss/

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*(4), 1523–1543. doi:10.1214/aos/1176350174

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Mahwah, NJ: Lawrence Erlbaum.

Hwang, W.-C., Wood, J. J., & Fujimoto, K. (2010). Acculturative family distancing (AFD) and depression in Chinese American families. *Journal of Consulting and Clinical Psychology, 78*(5), 655-667. doi:10.1037/a0020542

Ishwaran, H., & James, L. F. (2000). Approximate Dirichlet Process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, *11*(*3*), 508–532. doi:10.1198/106186002411

Ishwaran, H., and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173. doi:10.1198/016214501750332758

James, L. F., Lijoi, A., & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics, 36*(1), 76–97. doi:10.1111/j.1467-9469.2008.00609.x

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research, 34*(2), 245–268. doi:10.1207/S15327906Mb340205

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika, 68*(4), 563–583. doi:10.1007/BF02295612

Kalli, M., Griffin, J. E., & Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing, 21*(1), 93-105. doi:10.1007/s11222-009-9150-y

Kane, M. T. (2006). Validation (4th ed.). In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: Praeger.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*(7), 499–518. doi:10.1177/0146621608327800

Karabatsos, G., & Sheu, C. F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement, 28*(2), 110–125. doi:10.1177/0146621603260678

Karabatsos, G., & Walker, S. G. (2012a). Adaptive-modal Bayesian nonparametric regression. *Electronic Journal of Statistics, 6*, 2038–2068. Retrieved from http://projecteuclid.org.proxy.cc.uic.edu/ejs

Karabatsos, G., & Walker, S. G. (2012b). Bayesian nonparametric mixed random utility models. *Computational Statistics & Data Analysis*. doi:10.1016/j.csda.2011.10.014

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*(4), 345–355. doi:10.1177/014662169802200403

Kleinman, K. P., & Ibrahim, J. G. (1998a). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine, 17*(22), 2579–2596.

Kleinman, K. P., & Ibrahim, J. G. (1998b). A semiparametric Bayesian approach to the random effects model. *Biometrics, 54*(3), 921–938. doi:10.2307/2533846

Kottas, A., Müller, P., & Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics, 14*(3), 610–625. doi:10.1198/106186005X63185

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33*(2), 188–229. doi:10.1177/0049124103262065

Lijoi, A., Mena, R. H., & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association, 100*(472), 1278–1291. doi:10.1198/016214505000000132

Lijoi, A., Mena, R. H., & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(4), 715–740. doi:10.1111/j.1467-9868.2007.00609.x

Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*(4), 1223–1234.

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

MacEachern, S. N. (1999). Dependent nonparametric processes. In American Statistical Association (Ed.), *Proceedings of the Section on Bayesian Statistical Science* (pp. 50–55). Alexandria, VA: American Statistical Association.

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). Crete, Greece: International Society for Bayesian Analysis.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. doi:10.1007/bf02296272

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear modelling* (2nd ed.). New York, NY: Chapman and Hall.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–741. Retrieved from http://www.apa.org/pubs/journals/amp/

Miyazaki, K., & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika, 74*(3), 375–393. doi:10.1007/s11336-008-9108-6

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(4), 417–430. doi:10.1177/014662168200600404

Muliere, P., & Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, *26*(*2*), 283–297. doi: 10.2307/3315511

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*(1), 59–71. doi:10.1177/014662169001400106

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. doi:10.1177/014662169201600206

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36*(3), 217–232. doi:10.1111/j.1745-3984.1999.tb00555.x

Newton, M., Czado, C., & Chappell, R. (1996). Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association, 91*(433), 142–153. Retrieved from http://www.jstor.org/stable/2291390

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178. doi:10.3102/10769986024002146

Penfield, R. D. (2006, April). *A nonparametric method for assessing differential step functioning in polytomous items.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3), 187–210. doi:10.1111/j.1745-3984.2007.00034.x

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353–370. doi:10.1111/j.1745-3984.2003.tb01151.x

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*(1), 61–78. doi:10.1080/08957340802558367

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 125–167). New York, NY: Elsevier.

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice, 28*(1), 38–49. doi:10.1111/j.1745-3992.2009.01135.x

Penfield, R. D., Myers, N. D., & Wolfe, E. W. (2008). Methods for assessing item, step, and threshold invariance in polytomous items following the partial credit model. *Educational and Psychological Measurement, 68*(5), 717–733. doi:10.1177/0013164407312602

Poirier, D. J. (1998). Revising beliefs in nonidentified models. *Econometric Theory, 14*(4), 483-509. Retrieved from http://www.jstor.org/stable/3533214

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Regazzini, E., Lijoi, A., & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics, 31*(2), 560–585. Retrieved from http://www.jstor.org/stable/3448406

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 59*(4), 731-792. Retrieved from http://www.jstor.org/stable/2985194

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. New York, NY: Springer Verlag.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association, 96*(453), 20–31. Retrieved from http://www.jstor.org/stable/2670337

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12*(4), 397-409. doi:10.1177/014662168801200408

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282. doi:10.1177/014662169001400305

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75–92. Retrieved from http://www.bps.org.uk/publications/journals/journaltitles/bjmsp.cfm

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). New York, NY: Waxmann.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100. doi:10.1007/BF02290599

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. doi:http://www.jstor.org/stable/2958889

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. doi:10.1007/BF02294360

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica, 4*(2), 639–650. Retrieved from http://www3.stat.sinica.edu.tw/statistica/

Sijtsma, K., & Molenaar, I. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Mahwah, NJ: Lawrence Erlbaum.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 40*(2), 106–108. doi:10.2307/2684866

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 64*(4), 583–639. Retrieved from http://www.jstor.org/stable/3088806

Stark, S., Chernyshenko, O., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. doi:10.1037/0021-9010.91.6.1292

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44–47. Retrieved from http://www.jstor.org/stable/2984877

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118–128. doi:10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Mahwah, NJ: Lawrence Erlbaum.

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics and Data Analysis, 22*(5), 537–557. doi:10.1016/0167-9473(96)00004-7

van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*(2), 272–279. doi:10.1007/s11336-010-9147-7

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach* (Unpublished doctoral dissertation). Retrieved from http://etd.lib.fsu.edu/theses/available/etd-11012006-131824/

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91*(433), 217–221. Retrieved from http://www.jstor.org/stable/2291398

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

von Davier, M., & Rost, J. (Eds.). (1995). *Polytomous mixed Rasch models*. New York, NY: Springer.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*(6), 389–406. doi:10.1177/0146621604268734

Wagner-Menghin, M. M. (2007). The mixed-Rasch model: An example for analyzing the meanings of response latencies in a personality questionnaire. In E. V. Smith Jr. & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 103–120). Maple Grove, MN: JAM.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education, 72*(3), 221–261. doi:10.3200/JEXE.72.3.221-261

Wang, W.-C. (2007). Assessment of differential item functioning. *Journal of Applied Measurement, 9*(4), 387–408.

Wang, W.-C., & Su, Y.-H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*(6), 450–480. doi:10.1177/0146621604269792

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479–498. doi:10.1177/0146621603259902

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*(1), 22–42. doi:10.1177/0146621605279867

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis. Rasch Measurement*. Chicago: MESA Press.

Yamamoto, K. (1987). *A model that combines IRT and latent class models* (Unpublished doctoral dissertation). University of Illinois, Champaign-Urbana.

VITA

KEN AKIRA FUJIMOTO

EDUCATION

August 2014	Doctor of Philosophy Measurement, Evaluation, Statistics, and Assessment Department of Educational Psychology University of Illinois at Chicago
December 2004	Master of Fine Arts Creative Writing–Fiction Department of English University of Arizona
May 1997	Bachelor of Arts Department of Psychology California State University Long Beach

SELECTED AWARDS

- Oxford Bibliographies in Psychology Graduate Student Award (differential item functioning entry)
- University of Illinois at Chicago University Fellow
- Graduate Minority Fellowship, University of Arizona
- Career Opportunity in Research Fellowship, sponsored by the National Institute of Mental Health

JOURNAL PUBLICATIONS

1. Fujimoto, K. A. & Hwang, W.-C. (in press). Acculturative family distancing: Psychometric analysis with the extended two-tier IRT. *Psychological Assessment*.
2. Fujimoto, K. A. & Karabatsos, G. (2014). Dependent Dirichlet process rating model. *Applied Psychological Measurement*, *38*(3), 217–228.
3. Ji, P. & Fujimoto, K. A. (2013). Measuring heterosexual LGBT ally development: A Rasch analysis. *Journal of Homosexuality*, *60*(12), 1695–1725.
4. Colwell, N., Gordon, R. A., Fujimoto, K. A., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver Interaction Scale: Results from the Early

Childhood Longitudinal Study-Birth Cohort. *Early Childhood Research Quarterly*, *28*(2), 218–233.

5. Gordon, R. A., Fujimoto, K. A., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology*, *49*(1), 146–160.
6. Smith, E. V. & Fujimoto, K. A. (2011). MultRasch SAS code for the estimation of the multiplicative Rasch model parameters. *Applied Psychological Measurement*, *35*(6), 485–486.
7. Hwang, W.-C., Wood, J. J., & Fujimoto, K. A. (2010). Acculturative family distancing (AFD) and depression in Chinese American families. *Journal of Consulting and Clinical Psychology*, *78*(5), 655–677.
8. Finlayson, M. L., Peterson, E. W., Fujimoto, K. A., & Plow, M. A. (2009). Rasch validation of the Falls Prevention Strategies Survey. *Archives of Physical Medicine and Rehabilitation*, *90*(12), 2039–2046.
9. Reist, C., Mazzanti, C., Vu, R., Fujimoto, K. A., & Goldman, D. (2004). Inter-relationships of intermediate phenotypes for serotonin function, impulsivity, and a 5-HT2A candidate allele: His452Tyr. *Molecular Psychiatry*, *9*(9), 871–878.
10. Reist, C., Nakamura, K., Sagart, E., Sokolski, K. N., & Fujimoto, K. A. (2003). Impulsive aggressive behavior: Open-label treatment with citalopram. *The Journal of Clinical Psychiatry*, *64*(1), 81–85.
11. Duffy, J. G., Reist, C., Fujimoto, K. A., & Cahill, L. (2001). Beta-adrenergic blockade and emotional memory in PTSD. *International Journal of Neuropsychopharmacology*, *4*, 377–383.
12. Reist, C., Vu, R., Coccaro, E. F., & Fujimoto, K. A. (2000). Serotonin-stimulated calcium release is decreased in platelets from high impulsivity patients. *The International Journal of Neuropsychopharmacology*, *3*(4), 315–320.
13. Strybel, T. Z. & Fujimoto, K. A. (2000). Minimum audible angles in the horizontal and vertical planes: Effects of stimulus onset asynchrony and burst duration. *The Journal of the Acoustical Society of America*, *108*(6), 3092–3095.
14. Albers, L. J., Reist, C., Vu, R. L., Fujimoto, K. A., Ozdemir, V., Helmeste, D., Poland, R., & Tang, S. W. (2000). Effect of venlafaxine on imipramine metabolism. *Psychiatry Research*, *96*(3), 235–243.

OTHER PUBLICATIONS AND REPORTS

1. Bada, X., Schmit, J., & Fujimoto, K. A. (2011). Does birth place matter? Determinants of non-electoral civic and political engagement. *Dialogo*, *14*, 34–36.
2. Pallares, A., Guridy, V., & Fujimoto, K. A. (2011). Gender, education and civic engagement. *Dialogo*, *14*, 38–40.

3. De los Angeles Torres, M., Lesinski, N., & Fujimoto, K. A. (2011). Does age matter? *Dialogo*, *14*, 41–42.

4. Whalen, S. P., Fujimoto, K. A., & Xiong, Y. (2008). *Hours of OST participation and developmental benefits for student participants in the Chicago Public Schools Community Schools Initiative (CSI)*. A report of the UIC Community Schools Evaluation project.

CONFERENCE PRESENTATIONS

1. Fujimoto, K. A. & Karabatsos, G. (2014, July). *Multiple Dependent Dirichlet Process Rating Model*. Paper will be presented at the 79th Annual Meeting of the Psychometric Society, Madison, WI.

2. Fujimoto, K. A. (2014, April). *Bayesian extended two-tier full-information item factor analysis model*. Paper presented at the National Council on Measurement in Education, Philadelphia, PA.

3. Fujimoto, K. A. & Karabatsos, G. (2013, October). *The dependent Dirichlet process rating model for health outcomes*. Poster session presented at the 10th International Conference on Health Policy Statistics, Chicago, IL.

4. Fujimoto, K. A. & Karabatsos, G. (2013, April). *The dependent Dirichlet process rating model for DIF analysis*. Paper presented at the National Council on Measurement in Education, San Francisco, CA.

5. Colwell, N., Gordon, R. A., Fujimoto, K. A., Kaestner, R., & Korenman, S. (2013, April). *Domain-specific quality measures for early childhood programs: New evidence from the Study of Early Child Care and Youth Development*. Paper presented at the symposium, "New Insights into Early Care and Education Quality and Child Development: Profiles of Care and Domain-Specific Aspects of Quality" at the Society for Research in Child Development, Seattle, WA.

6. Gordon, R. A., Hofer, K., Fujimoto, K. A., Colwell, N., Kaestner, R., & Korenman, S. (2013, April). *Measuring aspects of child care quality specific to domains of child development: An indicator-level analysis of the ECERS-R*. Paper presented in the symposium, "Measuring Early Care and Education Quality: New Insights about the Early Childhood Environment System Rating Scale-Revised" at the Society for Research in Child Development, Seattle, WA.

7. Colwell, N., Gordon, R. A., Fujimoto, K. A., Kaestner, R., & Korenman, A. (2012, November). *New evidence on the validity of the Arnett Caregiver Interaction Scale: Results for preschoolers in the Early Childhood Longitudinal Study-Birth Cohort*. Poster session presented at the National Council of Family Relations annual conference, Phoenix, AZ.

8. Hwang, W.-C., Mak, E., Fujimoto, K. A., Li, R., Ng, W., Chiu, E., Butner, J., Myers, H. F., & Miranda, J. (2012, August). *Culturally adapting psychotherapy: Moving from frameworks to evidence-based clinical practice*. Paper presented at the American Psychological Association (APA) conference, Orlando, FL.

9. Fujimoto, K. A. (2012, April). *Random item Rasch model: Effectiveness in detecting differential item functioning through the variance component*. Paper presented at the International Objective Measurement Workshop, Vancouver, Canada.
10. Colwell, N., Gordon, R. A., & Fujimoto, K. A. (2012, April). *New evidence on the validity of the first grade classroom observation system*. Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.
11. Gordon, R. A., Fujimoto, K. A., Kaestner, R., Korenman, S., & Abner, K. (2010, June). *Psychometric properties of the ECERS-R in a national sample of four-year-olds*. Poster presented at the Institute of Education Sciences Research Conference, National Harbor, MD.
12. Finlayson, M., Peterson, E., Fujimoto, K. A., & Plow, M. (2009, May). *Rasch validation of the Falls Prevention Strategies Survey*. Paper presented at the consortium of Multiple Sclerosis Centers 23rd Annual Meeting, Atlanta, GA.
13. Goldman, S. R., Lawless, K. A., Gomez, K., Bertenthal, M., Braasch, J., MacLeod, S., Manning, F. H., Fujimoto, K. A., & Manderino, M. (2007, April). *Assessing multiple-source digital literacy skills*. Paper presented at the symposium 21st Century Literacy: A Symposium in Honor of Michael Pressley at the American Educational Research Association, Chicago, IL.

TEACHING EXPERIENCE

Summer 2011	Graduate Instructor, *Essentials of Quantitative Inquiry in Education* Department of Educational Psychology University of Illinois at Chicago
2007–2009	Graduate Assistant, Measurement, Evaluation, Statistics, and Assessment Laboratory University of Illinois at Chicago
Fall 2008	Teaching Assistant, *ANOVA and Regression* Department of Educational Psychology University of Illinois at Chicago

PROFESSIONAL SERVICE

- Ad hoc reviewer, *Psychometrika*
- Ad hoc reviewer, *Applied Psychological Measurement*
- Reviewer, National Council on Measurement in Education (2013)
- Ad hoc reviewer, *Canadian Journal of Occupational Therapy*

- Co-reviewer (with Rachel A. Gordon, Ph.D.), Society for Research in Child Development, Themed Meeting: Developmental Methodology (2011)

PROFESSIONAL MEMBERSHIPS

- American Educational Research Association
- National Council on Measurement in Education
- Psychometric Society