A Bayesian Beta-Mixture Model for Nonparametric IRT (BBM-IRT)

Ethan A. Arenson and George Karabatsos[1]

University of Illinois-Chicago

September 5, 2017

Introduction

Item Response Theory (IRT) is a successful enterprise that provides a class of useful statistical models for the analysis of item response data (Hambleton & Swaminathan, 1985; van der Linden, 2016).

Any IRT model posits a probabilistic relationship between each person's response to each test item, based on person ability and item parameters. To explain, let Y_{ni} be the item response random variable for person n and test item i, for persons and items indexed by $n = 1,\ldots,N$ and by $i = 1,\ldots,I$ (resp.). Nearly all IRT models make at least the following three assumptions (Junker & Sijtsma, 2001):

1) *Unidimensionality*: Person ability, θ, is real-valued (possibly multidimensional);

2) *Local Independence*: The conditional distribution of the responses to the test items satisfies:

$$\Pr(Y_1 = y_1,\ldots,Y_I = y_I \,|\theta) = \prod_{i=1}^{I} \Pr(Y_i = y_i \,|\theta). \quad (1)$$

3) *Monotonicity*: The Item Step Response Function (ISRF),

$$\Pr(Y_i \geq k \mid \theta) \text{ is monotone non-decreasing in } \theta, \text{ for } i = 1,\ldots,I \text{ and } k = 0,1,\ldots,K_i, \quad (2)$$

with $\Pr(Y_i = 1|\theta) = \Pr(Y_i \geq k|\theta)$ the Item Characteristic Curve (ICC) for a dichotomous ($K_i = 1$) items. Then, θ has positive correlation with the total test score ($\sum_{i=1}^{I} Y_i$) (Van der Ark & Bergsma, 2010).

These three assumptions exactly describe the nonparametric, monotone homogeneity (MH) IRT model (Mokken, 1971). It is the most general monotone IRT model which nests the 4-parameter logistic model (Hambleton & Swaminathan, 1985); the graded response logistic model (Samejima, 1969); and all other monotone IRT models (van der Ark, 2001) as special cases.

[1] Supported by NSF-MMS Research Grant SES-1156372, funded to Karabatsos, Principal Investigator.

In this article, we focus on unidimensional IRT for dichotomous items. While parametric IRT models provide a certain elegance and computational simplicity, nonparametric IRT models are more informative, and more closely describe the true item response functions that underlie real data. This contrasts with parametric IRT models which assume that ICCs follow a parametric distribution function, such as the logistic function (e.g., van der Linden, 2016). Also, a nonparametric IRT model can provide better fit to data compared to parametric IRT models, be used to evaluate the fit of the latter, and promote coherent statistical inference from a Bayesian perspective (Karabatsos & Walker, 2009a).

Several researchers have proposed various MH models defined by generalized linear models that specify the ICC as an inverse-link function parameter that is monotone in θ, and give support to the entire space of monotone cumulative distribution functions (c.d.f.s). Qin (1998) and Duncan and MacEachern (2008) proposed a Bayesian nonparametric (BNP) model that constructed monotone ICCs by a Dirichlet process centered on a 2-parameter logistic IRT model. Karabatsos (2017) modeled ICCs by a BNP infinite-mixture of normal c.d.f.s for the latent item response variables, with person- and item-dependent mixture weights. Karabatsos and Sheu (2004) and Tijmstra et al. (2013) proposed isotonic regression, using a Bayesian and frequentist approach (resp.), assuming discrete-valued θ. Luzardo and Rodriguez (2015) used classic nonparametric kernel regression methods to estimate monotone ICCs. Finally, Karabatsos and Walker (2009b, 2010) presented a BNP beta-mixture model for test score equating.

We propose a simple and flexible BNP IRT model for dichotomous items and continuous-valued ability (θ), extending a generalized linear model with unknown link function parameter (Mallick & Gelfand, 1994). Our BNP IRT model maps the unidimensional ability parameter θ from the real-line onto (0, 1), and constructs a (random) monotone ICC (inverse-link) by a flexible finite-mixture of beta c.d.f.s. In fact, any smooth c.d.f. on (0, 1) can be approximated arbitrarily-well by a suitable finite mixture of beta c.d.f.s (Diaconis & Ylvasiker, 1985).

The Bayesian beta-mixture IRT model (BBM-IRT) is more flexible than traditional parametric IRT models, which make logistic or normal distributional assumptions about the ICCs. The BBM-IRT model

allows one to estimate more accurately estimate ICCs which may have shapes that would be considered misfitting under the traditional models. Also, the BBM-IRT model is more parsimonious and computationally feasible than previous BNP IRT models which can employ thousands of parameters.

The BBM-IRT model is completed by the specification of a joint prior distribution for the person ability parameters and the item-level mixture weight parameters, and the number of mixture components. Our IRT model is a flexible but "approximate" BNP model because it makes use finite instead of infinite-mixtures for more computational tractability. A mixture of 3 to 4 beta distributions was believed to provide adequate modeling flexibility (Mallick & Gelfand, 1994). This article shows that a mixture of 10 beta distributions, per test item, can provide gains in data fit for IRT modeling.

Section 2 presents our BBM-IRT model, and statistics for assessing the model's goodness-of-predictive fit. A simple iterative Markov chain Monte Carlo (MCMC) algorithm (Appendix) can be used for estimating the posterior distribution of the model parameters and their functionals of interest. Section 3 illustrates our IRT model through the analysis of a 20-item math exam data set, from a 2015 Trends in International Mathematics and Science Study (TIMSS) assessment of 8th grade students. Section 4 discusses conclusions and possible directions for future research.

Methodology: BBM-IRT model

Let $Y_{ni} \in \{0,1\}$ be dichotomous item-response variable, for person n and test item i. For a matrix of realized item response data, $\mathbf{Y} = (\underline{y_{ni}})_{N\times I}$, our BBM-IRT model is defined by:

$$\begin{gathered}\Pr(Y_{ni}=1|\theta_n;\omega_i,\boldsymbol{\xi}) = \sum_{j=1}^{J}\omega_{ij}\mathrm{Binc}\left(\frac{\exp(\theta_n/2)}{1+\exp(\theta_n/2)};\,\xi_1 j,\,\xi_2\cdot(J-j+1)\right)\\ (\omega_{1i},\ldots,\omega_{Ji}) \sim \mathrm{Di}(\alpha_{1i},\ldots,\alpha_{Ji}),\ \text{for } i=1,\ldots,I,\\ \xi_1,\xi_2 \overset{iid}{\sim} \mathrm{U}(.01,J),\\ \theta_n \sim \mathrm{N}(0,1),\ \text{for } n=1,\ldots,N.\end{gathered} \tag{3}$$

Each monotone ICC $\Pr(Y_i = 1|\theta)$ is modeled by the incomplete beta function (Binc), with beta mixture weights (ω_{ji}) assigned a Dirichlet (Di) prior distribution (with a non-informative, uniform prior

defined by $\alpha_{ji} \equiv 1$, for $j = 1,\ldots,J$), and scaling beta-shape parameters (ξ_1, ξ_2) assigned a uniform U(.01, J) prior distribution. Each test item i has $J - 1$ mixture-weight parameters, with $\omega_{Ji} = 1 - \sum_{j=1}^{J-1} \omega_{ji}$. The N person ability parameters θ_n are assigned a standard normal N(0,1) prior distribution. The probability density functions (p.d.f.s) of these distributions (denoted $n(\cdot \mid 0,1)$, $di(\cdot \mid \alpha_1,\ldots,\alpha_J)$, $u(\cdot \mid 0,J)$) are defined in standard texts (Kotz et al. 2004; Johnson et al. 1994, 1995).

The specification of the BBM-IRT model (3) mainly requires the choice of the number of beta mixture components (J), which can be sufficiently large so that the beta mixture well-approximates the entire space of monotone ICCs. The term $\frac{\exp(\cdot/2)}{1+\exp(\cdot/2)}$ in (3) maps from the real line (the space of θ) onto (0,1), using a constant (e.g., 2) to bound the Binc function within (0,1) (Mallick & Gelfand, 1994).

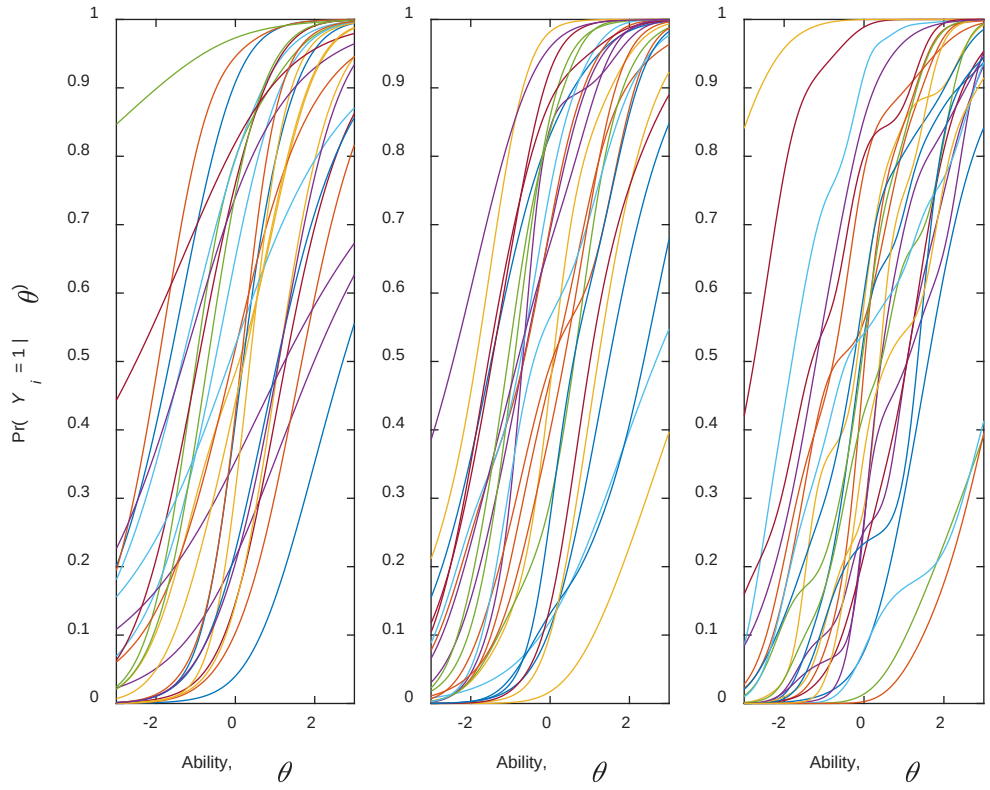

Figure 1. Random samples of 25 ICCs from the BBM-IRT model, where J = 3 (left panel), J = 5 (middle), and J = 10 (right).

Figure 1 displays three groups of samples of monotone ICCs. Each group was generated from 25 samples of the mixture weights ($\boldsymbol{\omega}_i = (\omega_{ji})_{j=1}^{J}$) from the uniform Dirichlet prior, and samples of ξ_1 and ξ_2 from the uniform U(.01, J) distribution (resp.), for J = 3 (left panel), J = 5 (middle), and J = 10 (right).

The ICCs in the left panel (J = 3) resemble the ICCs of a 2-parameter logistic (2PL) IRT model. If J = 1 and $\xi = \xi_1 = \xi_2$, then the BBM-IRT model reduces to a Rasch-type IRT model with common item discrimination parameter, ξ. The middle and right panels show that as J is increased, the ICCs become wigglier and more flexible. The right panel shows that J = 10 mixture components defines a BBM-IRT model that broadly and flexibly supports the entire space of monotone ICCs. The BBM-IRT model is thus a monotone IRT model, and a highly-parametric BNP model (Müller & Quintana, 2004).

For the BBM-IRT model (3), the joint posterior p.d.f. (distribution) of the model parameters, $\boldsymbol{\zeta} = ((\theta_n)_{n=1}^{N}, (\boldsymbol{\omega}_i)_{i=1}^{I}, \xi_1, \xi_2)$ is given by (up to a normalizing constant):

$$\begin{aligned}\pi(\boldsymbol{\zeta}|\mathbf{Y}) \propto & \prod_{n=1}^{N}\prod_{i=1}^{I}\{\Pr(Y_{ni}=1|\theta_n;\boldsymbol{\omega}_i,\xi)\}^{y_{ni}}\{1-\Pr(Y_{ni}=1|\theta_n;\boldsymbol{\omega}_i,\xi)\}^{1-y_{ni}} \\ & \times \prod_{n=1}^{N}\mathrm{n}(\theta_n|0,1)\prod_{i=1}^{I}\mathrm{di}(\boldsymbol{\omega}_i|\alpha_{1i},\ldots,\alpha_{Ji})\mathrm{u}(\xi_1|0,J)\mathrm{u}(\xi_2|0,J),\end{aligned} \tag{4}$$

with ICCs $\Pr(Y_{ni} = 1 \mid \theta_n; \boldsymbol{\omega}_i, \xi)$ defined by (3), and corresponding posterior c.d.f. $\Pi(\boldsymbol{\zeta} \mid \mathbf{Y})$. The model's posterior predictive expectation ($\mathbb{E}$) and variance ($\mathbb{V}$) of the item response variable Y_{ni} is given by (for all persons n = 1,…,N and all items i = 1,…,I):

$$\begin{aligned}\mathbb{E}_{NI}(Y_{ni}) &= \int \Pr(Y_{ni}=1|\theta_n;\boldsymbol{\omega}_i,\xi)\mathrm{d}\Pi(\boldsymbol{\zeta}|\mathbf{Y}), \\ \mathbb{V}_{NI}(Y_{ni}) &= \int \Pr(Y_{ni}=1|\theta_n;\boldsymbol{\omega}_i,\xi)\{1-\Pr(Y_{ni}=1|\theta_n;\boldsymbol{\omega}_i,\xi)\}\mathrm{d}\Pi(\boldsymbol{\zeta}|\mathbf{Y}).\end{aligned} \tag{5}$$

We estimate the BBM-IRT model's posterior distribution (4) and the posterior predictive quantities (5) with an adaptive random-walk Metropolis-Hastings MCMC algorithm. See the Appendix for details.

One may compare the predictive fit between BBM-IRT models to the data, which may differ by choice of J or prior distribution. For each model indexed by m = 1,…,M, the $D(m)$ criterion measures posterior predictive model fitness to the data $\mathbf{Y}$ (Laud & Ibrahim, 1995), and is defined by:

$$D(m) = \sum_{n=1}^{N} \sum_{i=1}^{I} \{y_{ni} - \mathbb{E}_{NI}(Y_{ni}|m)\}^2 + \mathbb{V}_{NI}(Y_{ni}|m). \quad (6)$$

The first term in (6) measures goodness-of-fit to the sample data (**Y**). The second term is a model complexity penalty. Among the *M* Bayesian models compared, the model with the best predictive utility for the given data set (**Y**) is identified as the model with the smallest value of *D*(*m*). The *D*(*m*) criterion is often used in Bayesian data analysis practice, and is easier to compute compared to other criteria.

The fit of a single BBM-IRT model (*m*) can be assessed by standardized item-response residuals:

$$z_{ni} = \frac{y_{ni} - \mathbb{E}_{NI}(Y_{ni}|m)}{\{\mathbb{V}_{NI}(Y_{ni}|m)\}^{1/2}}, \text{ for } n = 1,\ldots,N \text{ and } i = 1,\ldots,I \quad (7)$$

An absolute residual $|z_{ni}|$ exceeding 2 or 3 suggests that the response y_{ni} is an outlier under the model.

Results: TIMSS data analysis

We illustrate our BBM-IRT model through the analysis of 2015 TIMSS data on a basic math and algebra assessment of 716 American 8th grade students. The data set contains the students' individual responses to 20 math items, each response scored as correct ($Y_{ni} = 1$) or incorrect ($Y_{ni} = 0$). The data contains 639 unique item response patterns on the 20-item test. The Supplementary material of this article provides the TIMSS data set and the descriptions of the 20 items. It also provides the MATLAB code (Natick, VA) files that were used to run the MCMC sampling algorithm to analyze the TIMSS data using BBM-IRT and the 2PL IRT models, and to produce the results reported here.

We fit the BBM-IRT model to the TIMSS data using $J = 10$ components, a uniform Dirichlet prior for the mixture weights of the test items, and a N(0,1) prior for the 716 student math ability parameters (resp.). We estimated the posterior distribution of the model by running the MCMC algorithm (Appendix) for 100K sampling iterations.

The left panel of Figure 2 presents the marginal posterior means and ±2 standard deviations of the person ability parameter, for each of the 639 unique item response patterns (resp.). The right panel presents the marginal posterior mean estimates for the 20 ICCs of the test items (resp.). It shows multiple crossings

among the 20 ICCs, with some ICCs fluctuating more than others. These ICC results exhibit the flexibility and monotonicity of the BBM-IRT model, which may be misdiagnosed as outlying according to traditional IRT models that assume more restrictive logistic or normal ICCs.

Some of the estimated ICCs in Figure 2 have non-zero lower asymptotes, indicating the presence of guessing among low-ability examinees for these items. Thus, the BBM-IRT model (and its MCMC algorithm) can account for lucky-correct item responses among low-ability examinees. It does so while avoiding the issues of estimating the chance parameter in the three-parameter logistic IRT model, using either marginal maximum likelihood or Bayesian methods.

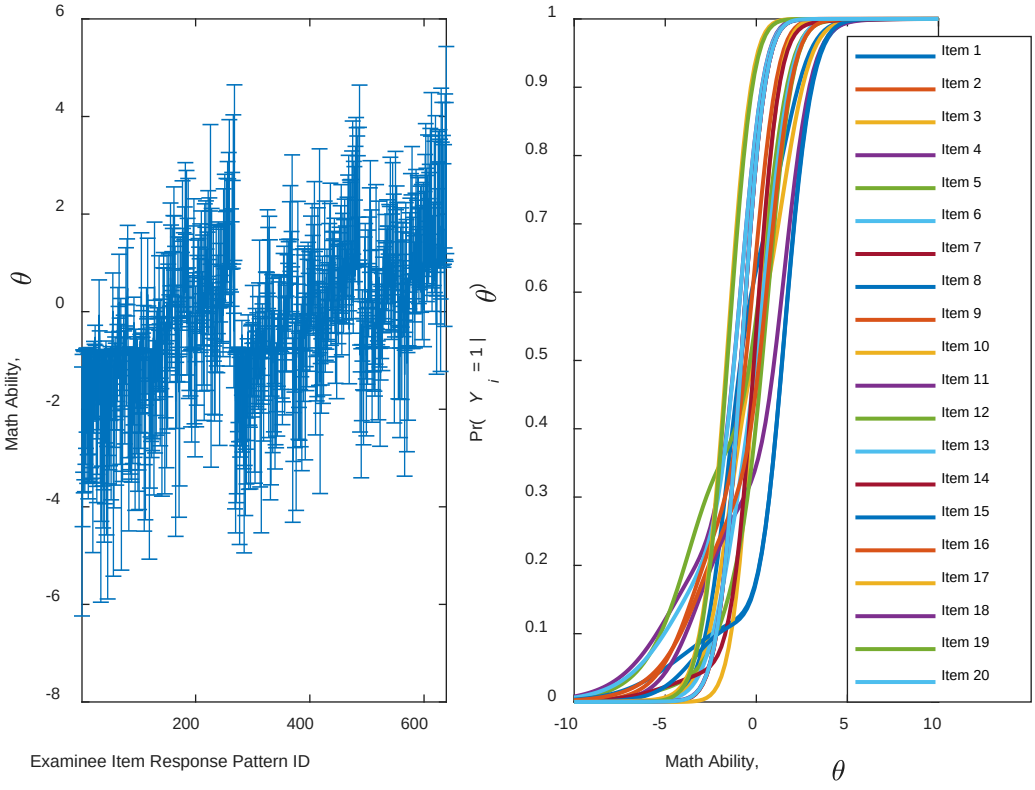

Figure 2. Left: Marginal posterior mean and +/–2 standard deviation for of the 639 unique item response patterns of the TIMSS exam. Right: Marginal posterior mean of the 20 TIMSS ICCs.

The following two diagnostic methods (Flegal & Jones, 2011) can be used to assess the convergence of the 100K MCMC parameter samples to the IRT model's exact posterior distribution. First, trace plots was used to evaluate the mixing (sampling independence) of the MCMC chain of each model

parameter, over the 100K iterations. Second, a non-overlapping batch means analysis of the chain was used to calculate the Monte Carlo 95% confidence interval half-width (95%MCCIhw) for each marginal posterior mean and posterior variance estimate. MCMC convergence can always be improved by running the MCMC chain beyond 100K sampling iterations.

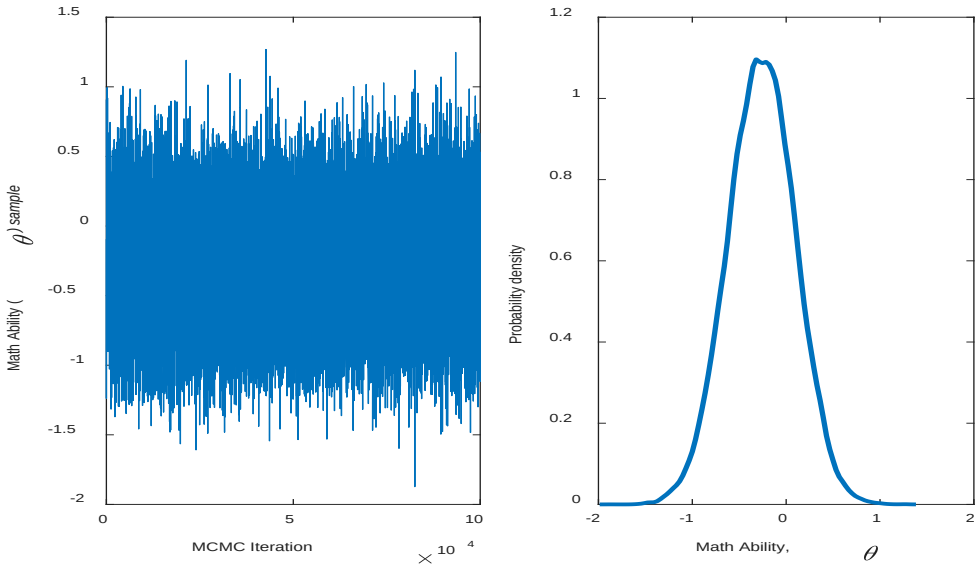

Figure 3. Trace and density plot of the math ability (θ) of 500th unique examinee item-response pattern.

Figure 3 illustrates these two MCMC convergence diagnostic methods for $\theta_{(500)}$, the ability parameter corresponding to the 500th unique item response pattern in the TIMSS data. The left panel of Figure 3 presents the trace plot of $\theta_{(500)}$ over the 100K MCMC sampling iterations. This plot supports good mixing and (near) sampling independence of the chain. The right panel presents a kernel density estimate of the marginal posterior distribution of $\theta_{(500)}$. The marginal posterior mean and variance of $\theta_{(500)}$ is –.3 and .1, with corresponding 95%MCCIhw values of .00 and .00. It can be concluded that the MCMC samples of the parameter $\theta_{(500)}$ have adequately converged to its marginal posterior distribution. Similar conclusions about MCMC convergence can be reached about the ability parameter for each of the 639 unique item response patterns, and about points on the ICC curve, for each individual test item.

For comparison, several different versions of the BNP-IRT model were fit to the TIMSS data. They employed either J = 3, 5, and 10 mixture components; either a standard normal N(0,1) prior distribution, a

left-skewed normal mixture prior .25×N(−1,1) + .75×N(1,1), or a right-skewed normal mixture prior .75×N(−1,1) + .25×N(1,1) on the ability parameters; and a uniform Dirichlet prior. The Bayesian 2PL model was also fit to the data, defined by:

$$\Pr(Y_{ni} = 1 \mid \theta_n, \alpha_i, \beta_i) = 1/[1+\exp(-\alpha_i\theta_n + \beta_i)], \quad \text{for } n = 1,\ldots,N, \text{ and } i = 1,\ldots,I, \tag{8}$$

with a N(0,1) prior for the person ability parameters (θ_n), a N(0,4) prior for the item difficulty parameters (β_i), and a N(0,1/4) prior for the log slope log(α_i) parameters (resp.), suggested for analyzing data from large scale testing (Patz & Junker, 1999, p.163). The 2PL model was fit using an adaptive version of a published random-walk Metropolis MCMC algorithm (Patz & Junker, 1999). To estimate the posterior distribution of each of these compared IRT models, the MCMC algorithm was run for 100K iterations. In each case, MCMC convergence analyses can be shown to yield similar results as before.

		Proportion residuals	
Bayesian IRT Model	$D(m)$	$\lvert z_{ni} \rvert \geq 2$	$\lvert z_{ni} \rvert \geq 3$
BBM-IRT, $J = 3$	4752	0.03	0.005
BBM-IRT, $J = 5$	4643	0.03	0.006
BBM-IRT, $J = 10$	4625	0.03	0.004
BBM-IRT, $J = 3$, left-skewed θ prior	4729	0.03	0.004
BBM-IRT, $J = 5$, left-skewed θ prior	4666	0.03	0.005
BBM-IRT, $J = 10$, left-skewed θ prior	4637	0.02	0.004
BBM-IRT, $J = 3$, right-skewed θ prior	4702	0.03	0.004
BBM-IRT, $J = 5$, right-skewed θ prior	4653	0.03	0.005
BBM-IRT, $J = 10$, right-skewed θ prior	4640	0.02	0.004
2PL Model	4717	0.03	0.01

Table 1. A comparison of the predictive fit between different IRT models.

For each IRT model, Table 1 summarizes the posterior predictive model fit statistic, $D(m)$, and the proportion of posterior predictive standardized residuals (z_{ni}) greater than 2 (and 3) in absolute magnitude. By considering both criteria, we find that the BBM-IRT model with $J = 10$ mixture components obtained the best predictive fit among all the IRT models compared, including the 2PL model which had about twice the number of outliers with residuals $\lvert z_{ni} \rvert \geq 3$. In terms of $D(m)$, the BBM-IRT model fit best under

symmetric priors for the ability parameters.

Figure 4 compares the marginal posterior mean estimates for the 20 ICCs, between the BBM-IRT models with J = 3, 5, and 10 components and the Bayesian 2PL model (resp.), and a N(0,1) prior for the ability parameters. Compared to the ICCs of the 2PL model, the ICCs of the 3-component BBM-IRT model were most similar, and the ICCs of the 10-component BBM-IRT model were most dissimilar.

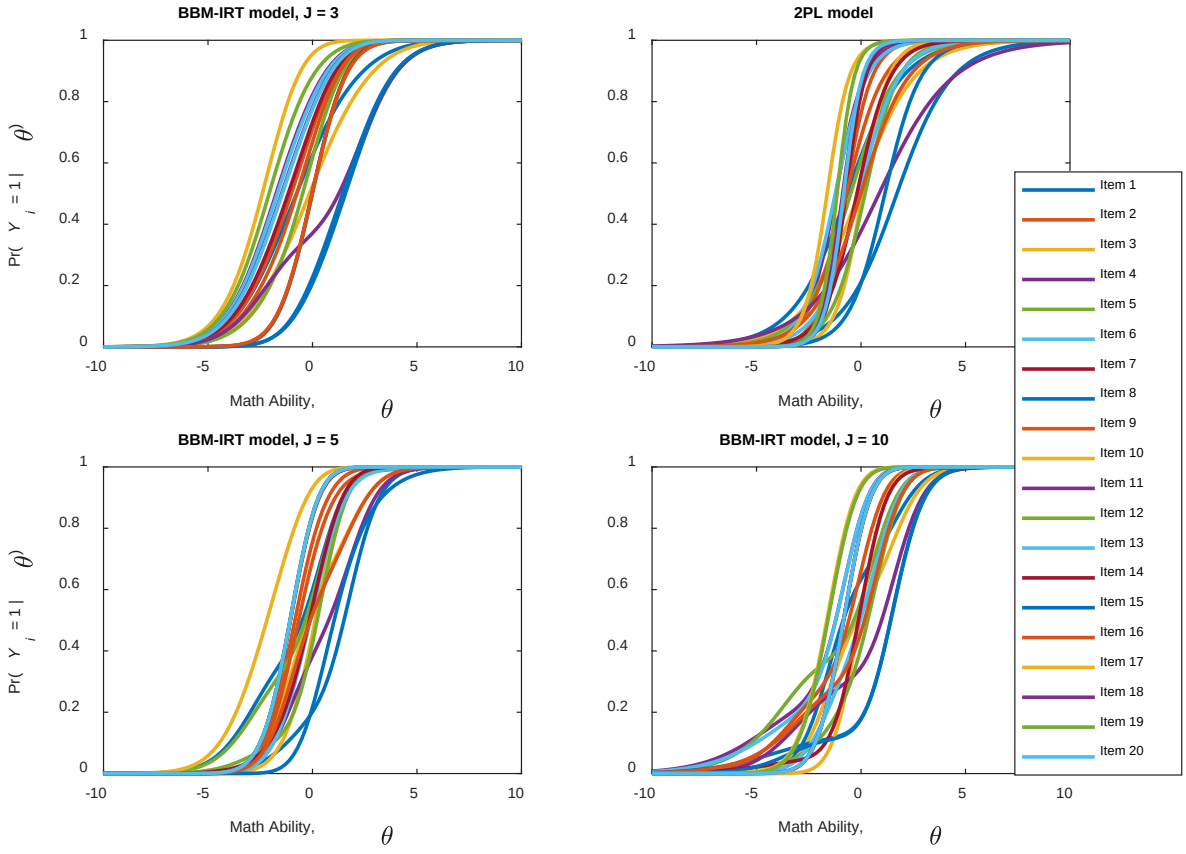

Figure 4. Estimated TIMSS ICCs for different IRT models.

Conclusions

We introduced a novel monotone, BBM-IRT model for dichotomous item responses and unidimensional ability. It provides a useful compromise between more restrictive parametric IRT models and more flexible and computationally intensive BNP models. The BBM-IRT flexibly models each ICC by a finite mixture of beta c.d.f.s, which approximately support the entire space of monotone ICCs. Posterior

inference of this model is possible through the application of a simple adaptive Metropolis MCMC algorithm. The usefulness of the BBM-IRT model was illustrated through the analysis of item response data from a TIMSS math assessment.

The BBM-IRT model shows promise for future research opportunities. For instance, one can extend this model to handle the analysis of polytomous item response data, by coding each observed polytomous response to a set of binary codes (Begg & Gray, 1984), or by replacing the Bernoulli kernel (incomplete Beta function) with a binomial or multinomial kernel in (3). In addition one can extend this model to handle multidimensional ability by assigning each person separate ability parameters for different subgroups of test items that measures different traits.

References

Atchadé, Y., & Rosenthal, J. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, *11*, 815-828.

Begg, C., & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, *71*, 11-18.

Diaconis, P., & Ylvisaker, D. (1985). Conjugate priors for exponential families. *Annals of Statistics*, *7*, 269-281.

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

Flegal, J., & Jones, G. (2011). Implementing Markov chain Monte Carlo: Estimating with confidence. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (p. 175-197). Boca Raton, FL: Chapman and Hall/CRC.

Hambleton, R., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. New York: Kluwer-Nijhoff.

Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions* (*vol.1*). New York: Wiley.

Johnson, N., Kotz, S., & Balakrishnan, N. (1995). *Continuous Univariate Distributions* (*vol.2*). New York: Wiley.

Junker, B., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65-81.

Karabatsos, G. (2017). Bayesian nonparametric IRT. In chapter 19, *Handbook of Item Response Theory: Models, Statistical Tools, and Applications, Volume 1* (W.J. van der Linden, Eds.). New York: Taylor & Francis.

Karabatsos, G., & Sheu, C.-F. (2004). Order constrained Bayes inference for dichotomous models of unidimensional non-parametric item response theory. *Applied Psychological Measurement*, *28*,110-125.

Karabatsos, G., & Walker, S.G. (2009a). Coherent psychometric modeling with Bayesian nonparametrics. *British Journal of Mathematical and Statistical Psychology*, *62*, 1-20.

Karabatsos, G., & Walker, S.G. (2009b). A Bayesian nonparametric approach to test equating. *Psychometrika*, *74*, 211-232.

Karabatsos, G., & Walker, S.G. (2010). A Bayesian nonparametric model for test equating. In Chapter 11 (pp. 175-184) of A. von Davier (Ed.), Statistical Models for Equating, Scaling, and Linking. New York: Springer.

Kotz, S., Balakrishnan, N., & Johnson, N. (2004). *Continuous Multivariate Distributions, Models and Applications*. New York: John Wiley and Sons.

Laud, P., & Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society*, *Series B*, *57*, 247-262.

Luzardo, M., & Rodriguez, P. (2015). A nonparametric estimator of a monotone item characteristic curve. In A. van der Ark, D. Bolt, W.-C. Wang, A. Douglas, & S.-M. Chow (Eds.), *Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society*, Madison, Wisconsin, 2014 (p. 99-108). Cham, Switzerland: Springer International Publishing.

Mallick, B., & Gelfand, A. (1994). Generalized linear models with unknown link functions. *Biometrika*, *81*, 237-245.

Mokken, R. (1971). *A Theory and Procedure of Scale Analysis*. The Hague: Mouton.

Müller, P., & Quintana, F.A. (2004). Nonparametric data analysis. *Statistical Science, 19*, 95-110.

Patz, R., & Junker, B. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response theory models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Qin, L. (1998). *Nonparametric Bayesian Models for Item Response Data* (Ph.D. Thesis). Unpublished doctoral dissertation, The Ohio State University.

Roberts, G., & Rosenthal, J. (2001). Optimal scaling of various Metropolis-Hastings algorithms. *Statistical Science, 16*, 351-367.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 18.

Tijmstra, J., Hessen, D., van der Heijden, P., & Sijtsma, K. (2013). Testing manifest monotonicity using order-constrained statistical inference. *Psychometrika, 78*, 83-97.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*, 273-282.

Van der Ark, L. A., & Bergsma, W. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*, 272-279.

van der Linden, W.J. (2016). *Handbook of Item Response Theory, Volumes 1, 2, 3*. Boca Raton, FL: Chapman & Hall/CRC.

Appendix: MCMC Algorithm for BBM-IRT Model

For the BBM-IRT model (3), a 3-step MCMC iterative sampling algorithm is used to estimate the posterior distribution (4) (density) of the N ability parameters and the I ICC parameters. A large number of sampling iterations (S) is run until the algorithm yields a sample that converges (approximately) to a sample from the posterior distribution (MCMC convergence).

The algorithm is initiated at stage $s = 0$, with model parameter values $(\theta_n^{(0)} \equiv 0)_{n=1}^{N}, (\omega_{ji} \equiv 1/J)_{j=1}^{J}$, $\xi_1 \equiv \xi_2 \equiv 1$, and their proposal variances, $(\tau_{\theta c}^{(0)} \equiv .01)_{n=1}^{N}, (\tau_{\omega i}^{(0)} \equiv .01)_{i=1}^{I}$, and $(\tau_{\xi}^{(0)} \equiv .01)_{i=1}^{I}$. Also, to speed up computations, the original item response data set $\mathbf{Y} = (y_{ni})_{N \times I}$ is collapsed into a smaller data set, $\mathbf{Y}_C = (y_{ci})_{C \times I}$, consisting of $C \leq N$ unique values of the item response vectors $(\mathbf{y}_c = (y_{c1}, \ldots, y_{cI}))_{c=1}^{C}$, with frequency counts $n_1, \ldots, n_c, \ldots, n_C$ (resp.).

Each iteration $s \in \{1, \ldots, S\}$ of the MCMC sampling algorithm runs the following three adaptive Metropolis sampling steps, applying the established methodology of Atchadé and Rosenthal (2005).

In Step 1, for each $c = 1, \ldots, C$ independently (concurrently), a candidate ability parameter $\theta_c^{*[s]}$ is drawn from the normal $\mathrm{N}(\theta_c^{[s]}, \tau_{\theta c}^{[s]})$ proposal distribution. The update $\theta_c^{[s]} \equiv \theta_c^{*[s]}$ is accepted with probability $P_{\theta c}^{[s]} = \min\{1, \rho_{\theta c}^{[s]}\}$, and otherwise $\theta_c^{[s]} \equiv \theta_c^{[s-1]}$ is set with probability $1 - P_{\theta c}^{[s]}$, where:

$$\rho_{\theta c}^{[s]} = \frac{\prod_{i=1}^{I} \{\Pr(Y_{ci} = 1 | \theta_c^{*[s]}; \boldsymbol{\omega}_i^{[s-1]}, \boldsymbol{\xi}^{[s-1]})\}^{y_{ui}} [1 - \Pr(Y_{ci} = 1 | \theta_c^{*[s]}; \boldsymbol{\omega}_i^{[s-1]}, \boldsymbol{\xi}^{[s-1]})]^{1-y_{ui}} \exp[-\tfrac{1}{2}(\theta_c^{*[s]})^2]}{\prod_{i=1}^{I} \{\Pr(Y_{ci} = 1 | \theta_c^{[s-1]}; \boldsymbol{\omega}_i^{[s-1]}, \boldsymbol{\xi}^{[s-1]})\}^{y_{ui}} [1 - \Pr(Y_{ci} = 1 | \theta_c^{[s-1]}; \boldsymbol{\omega}_i^{[s-1]}, \boldsymbol{\xi}^{[s-1]})]^{1-y_{ui}} \exp[-\tfrac{1}{2}(\theta_c^{[s-1]})^2]}. \quad (8)$$

In (8), a few terms cancel out, including the counts $(n_c)_{c=1}^{C}$, the constants of the normal proposal p.d.f.s, and, the normal proposal p.d.f.s due to their symmetry. Then the proposal variances are updated by $\tau_{\theta c}^{[s]} = \min[10^{-3}, \tau_{\theta c}^{[s-1]} + (1/s)(\rho_{\theta c}^{[s]} - .44)]$, towards achieving the optimal acceptance rate of .44 (see Roberts & Rosenthal, 2001) over the S MCMC iterations, for $c = 1, \ldots, C$.

In Step 2, for $i = 1, \ldots, I$ independently (concurrently), a candidate (log) mixture weight

parameter $\log \boldsymbol{\omega}_i^{*[s]} = \log(\omega_{1i}^{*[s]}, \ldots, \omega_{Ji}^{*[s]})$ is drawn from the *J*-variate normal proposal distribution, $\mathrm{N}(\log(\omega_{1i}^{*[s-1]}, \ldots, \omega_{Ji}^{*[s-1]}), \tau_{\omega i}^{[s-1]} \mathbf{I}_J)$. The update $\boldsymbol{\omega}_i^{[s]} \equiv \exp(\log \boldsymbol{\omega}_i^{*[s]})$ is accepted with probability $P_{\omega i}^{[s]} = \min\left\{1, \rho_{\omega i}^{[s]}\right\}$, and otherwise $\boldsymbol{\omega}_i^{[s]} \equiv \boldsymbol{\omega}_i^{[s-1]}$ is set with probability $1 - P_{\omega i}^{[s]}$, where:

$$\rho_{\omega i}^{[s]} = \frac{\prod_{c=1}^{C} n_c \{\Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{*[s]}, \boldsymbol{\xi}^{[s-1]})\}^{y_{ui}} [1 - \Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{*[s]}, \boldsymbol{\xi}^{[s-1]})]^{1-y_{ui}} \prod_{j=1}^{J} (\boldsymbol{\omega}_{ji}^{*[s]})^{\alpha_j - 1}}{\prod_{c=1}^{C} n_c \{\Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{[s-1]}, \boldsymbol{\xi}^{[s-1]})\}^{y_{ui}} [1 - \Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{[s-1]}, \boldsymbol{\xi}^{[s-1]})]^{1-y_{ui}} \prod_{j=1}^{J} (\boldsymbol{\omega}_{ji}^{[s-1]})^{\alpha_j - 1}}. \quad (9)$$

In (9), the Dirichlet proposal p.d.f. constants and the normal proposal p.d.f. cancel out. Then, the proposal variances are updated by $\tau_{\omega i}^{[s]} = \min[10^{-3}, \tau_{\omega i}^{[s-1]} + (1/s)(\rho_{\omega i}^{[s]} - .234)]$, towards achieving the optimal acceptance rate of .234 for multidimensional parameters (Roberts & Rosenthal, 2001), for $i = 1, \ldots, I$.

In Step 3, a candidate $\log(\xi_1, \xi_2)^{*[s]}$ is drawn from the bivariate normal proposal distribution, $\mathrm{N}(\log(\xi_1, \xi_2)^{[s-1]}, \tau_{\xi}^{[s-1]} \mathbf{I}_J)$. Thee update $(\xi_1, \xi_2)^{[s]} \equiv \exp(\log(\xi_1, \xi_2)^{*[s]})$ is accepted with probability $P_{\xi}^{[s]} = \min\left\{1, \rho_{\xi}^{[s]}\right\}$, and the old value is kept $(\xi_1, \xi_2)^{[s]} \equiv (\xi_1, \xi_2)^{[s-1]}$ with probability $1 - P_{\xi}^{[s]}$, where:

$$\rho_{\xi}^{[s]} = \frac{\prod_{c=1}^{C} \prod_{i=1}^{I} n_c \{\Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{*[s]}, \boldsymbol{\xi}^{*[s]})\}^{y_{ui}} [1 - \Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{*[s]}, \boldsymbol{\xi}^{*[s]})]^{1-y_{ui}} \mathbf{1}(.01 < \xi_1^{*[s]}, \xi_2^{*[s]} < J)}{\prod_{c=1}^{C} \prod_{i=1}^{I} n_c \{\Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{[s]}, \boldsymbol{\xi}^{[s-1]})\}^{y_{ui}} [1 - \Pr(Y_{ci} = 1 | \theta_c^{[s]}; \boldsymbol{\omega}_i^{[s]}, \boldsymbol{\xi}^{[s-1]})]^{1-y_{ui}} \mathbf{1}(.01 < \xi_1^{[s-1]}, \xi_2^{[s-1]} < J)}. \quad (10)$$

The uniform prior p.d.f. constants and the normal proposal p.d.f. cancel out of the ratio (10). Then, the proposal variance is updated by $\tau_{\xi}^{[s]} = \min[10^{-3}, \tau_{\xi}^{[s-1]} + (1/s)(\rho_{\xi}^{[s]} - .44)]$, towards achieving the optimal acceptance rate of roughly .234 for two-dimensional parameters (Roberts & Rosenthal, 2001).

As a simple by-product of the three-step MCMC algorithm, it is possible to estimate the marginal posterior average and variance of each ability parameter, θ_c (for $c = 1, \ldots, C$), and the marginal posterior mean and variance of each ICC, $\Pr(Y_i = 1 | \theta; \boldsymbol{\omega}_i, \boldsymbol{\xi})$, over a chosen fine grid of θ values. Specifically, at each MCMC iteration $s \in \{1, \ldots, S\}$, the marginal posterior expectation ($\mathbb{E}$) and variance ($\mathbb{V}$) of each θ_c is updated through Rao-Blackwellization, via the calculations:

$$\hat{\mathbb{E}}_{NI}^{[s]}(\theta_c) = [\theta_c^{[s]} + (s-1)\hat{\mathbb{E}}_{NI}^{[s-1]}(\theta_c)]/s;$$
$$\hat{\mathbb{V}}_{NI}^{[s]}(\theta_c) = [(\theta_c^{[s]})^2 + (s-1)\hat{\mathbb{E}}_{NI}^{[s-1]}(\theta_c^2)]/s. \quad (11)$$

The marginal posterior mean and variance of the ICC, $\Pr(Y_{ci} = 1|\boldsymbol{\theta};\boldsymbol{\omega}_i,\boldsymbol{\xi})$, at each chosen grid point θ_c and test item $i = 1,\ldots,I$, are updated by:

$$\hat{\mathbb{E}}_{NI}^{[s]}(Y_{ci}\,|\,m) = [\Pr(Y_i = 1\,|\,\theta_c;\boldsymbol{\omega}_i^{[s]},\boldsymbol{\xi}^{[s]}) + (s-1)\hat{\mathbb{E}}_{NI}^{[s-1]}]/s;$$
$$\hat{\mathbb{V}}_{NI}^{[s]}(Y_{ci}\,|\,m) = [\Pr(Y_i = 1\,|\,\theta_c;\boldsymbol{\omega}_i^{[s]},\boldsymbol{\xi}^{[s]})[1-\Pr(Y_i = 1\,|\,\theta_c;\boldsymbol{\omega}_i^{[s]},\boldsymbol{\xi}^{[s]})] + (s-1)\{\hat{\mathbb{V}}_{NI}^{[s-1]}(Y_{ci})\}]/s. \quad (12)$$

Similar methods compute the updates $\hat{\mathbb{E}}_{NI}^{[s]}(Y_{ni}\,|\,m)$ and $\hat{\mathbb{V}}_{NI}^{[s]}(Y_{ni}\,|\,m)$, for each person $n = 1,\ldots,N$ and item $i = 1,\ldots,I$. The updated estimate of the standardized item-response fit residual, per unique item response pattern, is given by $\hat{z}_{ci}^{[s]} = \{y_{ci} - \hat{\mathbb{E}}_{NI}^{[s]}(Y_{ci}\,|\,m)\}/\{\hat{\mathbb{V}}_{NI}^{[s]}(Y_{ci}\,|\,m)\}^{1/2}$, for $c = 1,\ldots,C$ and $i = 1,\ldots,I$. The updated estimate of the $D(m)$ model fit criterion is given by:

$$\hat{D}^{[s]}(m) = \sum_{c=1}^{C}\sum_{i=1}^{I} n_c[\{y_{ci} - \hat{\mathbb{E}}_{NI}^{[s]}(Y_{ci}\,|\,m)\}^2 + \hat{\mathbb{V}}_{NI}^{[s]}(Y_{ci}\,|\,m)]. \quad (13)$$