Educational and Psychological
Measurement
2016, Vol. 76(4) 662–684
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164415608418
epm.sagepub.com
SAGE

Rasch Model Parameter Estimation in the Presence of a Nonnormal Latent Trait Using a Nonparametric Bayesian Approach

Holmes Finch[1] and Julianne M. Edwards[1]

Abstract

Standard approaches for estimating item response theory (IRT) model parameters generally work under the assumption that the latent trait being measured by a set of items follows the normal distribution. Estimation of IRT parameters in the presence of nonnormal latent traits has been shown to generate biased person and item parameter estimates. A number of methods, including Ramsay curve item response theory, have been developed to reduce such bias, and have been shown to work well for relatively large samples and long assessments. An alternative approach to the nonnormal latent trait and IRT parameter estimation problem, nonparametric Bayesian estimation approach, has recently been introduced into the literature. Very early work with this method has shown that it could be an excellent option for use when fitting the Rasch model when assumptions cannot be made about the distribution of the model parameters. The current simulation study was designed to extend research in this area by expanding the simulation conditions under which it is examined and to compare the nonparametric Bayesian estimation approach to the Ramsay curve item response theory, marginal maximum likelihood, maximum a posteriori, and the Bayesian Markov chain Monte Carlo estimation method. Results of the current study support that the nonparametric Bayesian estimation approach may be a preferred option when fitting a Rasch model in the presence of nonnormal latent traits and item difficulties, as it proved to be most accurate in virtually all scenarios that were simulated in this study.

[1] Ball State University, Muncie, IN, USA

Corresponding Author:
Holmes Finch, Department of Educational Psychology, Ball State University, Muncie, IN 47306, USA.
Email: whfinch@bsu.edu

Keywords

item response theory, Rasch model, nonnormal latent trait, nonparametric Bayes, parameter estimation

Item response theory (IRT) is a mainstay in psychometrics and educational measurement. It provides researchers with information regarding performance of both items on an assessment and performance of individuals who take the assessment. Information about the items can then be used to refine the instrument and to select sets of items that maximize the information provided about individuals. In turn, information about individuals can be used to identify those in need of special psychological or educational services, for example. There exist a number of IRT models for dichotomous items, including the Rasch model, which is perhaps the simplest of these. For a dichotomously coded item, the Rasch model takes the form:

$$P\left(x_{ij}=1|\theta_i,b_j\right)=\frac{e^{\left(\theta_i-b_j\right)}}{1+e^{\left(\theta_i-b_j\right)}}, \tag{1}$$

where x_{ij} is the response to item j for individual i, where 1 is coded as endorsement (or correct) and 0 as nonendorsement; θ_i is the person-level latent trait being measured by the scale; and $b_j$is the location (or difficulty) for item j.

It should be noted that in the context of educational and cognitive assessments, b_j is frequently referred to as the item difficulty parameter. In other contexts, this parameter reflects the location of the item on the latent trait scale, but does not indicate difficulty in the standard educational sense. In this article, we will use the two terms interchangeably.

Of key interest in the current study is the fact that standard approaches for estimating the model in (1), such as marginal maximum likelihood (MML), joint maximum likelihood (JML), and maximum a posteriori (MAP) assume that θ is normally distributed (de Ayala, 2009). A number of researchers have examined the impact of nonnormally distributed latent traits in the population on the estimation of both person and item parameters in the context of standard IRT, including the Rasch model for dichotomous data. Monte Carlo simulation work in this area has demonstrated repeatedly that when the latent trait is skewed, estimates of both b_j and θ based on MML are deleteriously affected (e.g., Stone, 1992; Woods, 2006, 2008; Woods & Linn, 2009; Woods & Thissen, 2006). Collectively this earlier work also demonstrated that the degree of bias in these parameter estimates was concomitant with the degree of skewness in θ, such that more skewness led to more biased estimates. This general finding of poor performance in the presence of skewed latent traits does appear to be mitigated, to some extent, by characteristics of the data itself. For example, Stone (1992) and Seong (1990) both found that bias in item parameter estimates was lower for longer instruments (i.e., more items) and for larger sample sizes, when expected a posteriori (EAP) estimation was used to estimate the latent trait. In addition, when

MAP was employed the impact of skewness on the latent trait was largely overcome for samples of 1,000 examinees and instruments of 40 items or more (Kirisci, Hsu, & Yu, 2001). In summary, prior research has generally found that when the latent trait is skewed, recovery of person parameter estimates will be biased, as compared to when the latent trait is normally distributed, for the most commonly used methods of estimation, MML, EAP, and MAP (e.g., Kirisci & Hsu, 1995; Stone, 1992; Woods, 2007, 2008).

Extending this work to include the distribution of item difficulty parameters for all items on a scale, Sass, Schmitt, and Walker (2008) examined the interaction between the distribution of θ and the collective distribution of item difficulty parameters. These authors were particularly interested in characterizing item and person parameter estimation bias when each of four approaches for estimating θ were used: MML, MAP with a normal prior, EAP using a prior distribution matching the distribution used to simulate the data, and EAP using empirical weights to characterize the latent trait distribution. Sass et al. simulated θ to be either normal, or with a skewness of 1.6 or −1.6. In addition to manipulating the latent trait distribution, these authors also manipulated the item difficulty parameters into three conditions that they referred to as normal, negatively skewed, and positively skewed. The values of the item difficulty values under each condition appear in Table 1, and will be used in the current study. We elected to use the same values as those in Sass et al. because we wanted to make the results of the current study, particularly with respect to a relatively new estimation method based on a nonparametric Bayesian approach, as directly comparable to those in prior research as possible. We do acknowledge, however, that it would also have been reasonable to generate new item parameter values from the same distribution as those used in the earlier study, rather than to use the actual values themselves as we have done. It is also important to note that whereas the current study used the Rasch model to generate and analyze item response data, Sass et al. used the 2-parameter logistic (2PL) model for this purpose.

Sass et al. (2008) were particularly interested in the combined impact of the collective item difficulty distribution and the latent trait distribution on the estimation of IRT model parameters. Their results reinforced prior work, showing that model parameter bias was lower when θ was normally distributed, and when the item difficulty values conformed to the normal pattern in Table 1. The poorest performance in terms of both item and person parameter estimation was associated with positive skew of both the items and θ. The authors hypothesized that this result might be due to a relative lack of items available for estimating θ at the highest end of the distribution, thereby creating a cascade of estimation problems across all levels of the latent trait (Sass et al., 2008). Finally, of the four estimation methods that were studied, EAP with the prior distribution matching the data generating distribution was found to perform the best, which was not surprising given the use of the correct prior. However, as Sass et al. (2008) noted, this is not a realistic scenario because in most cases researchers will not know what the population distribution of θ actually is. Other than EAP with the correct prior, MAP and MML both performed comparably

Table 1. Data Generating Item Difficulty Values.

	Difficulty		
Item number	Normal	Negative skew	Positive skew
1	1.08	−0.49	0.49
2	−0.01	2.06	−2.06
3	0.73	2.12	−2.12
4	−0.03	1.37	−1.37
5	1.02	1.29	−1.29
6	−0.15	1.96	−1.96
7	−2.24	1.67	−1.67
8	−1.98	1.16	−1.16
9	−0.24	2.30	−2.30
10	0.31	−0.51	0.51
11	0.43	−0.01	0.01
12	−0.71	1.27	−1.27
13	1.66	−0.11	0.11
14	0.26	−0.62	0.62
15	0.14	1.77	−1.77
16	−2.26	1.15	−1.15
17	−0.01	1.14	−1.14
18	0.05	1.97	−1.97
19	−0.12	1.72	−1.72
20	−0.39	2.16	−2.16
21	−0.26	0.39	−0.39
22	0.20	0.39	−0.39
23	0.59	1.95	−1.95
24	3.11	1.79	−1.79
25	−0.58	−0.87	0.87
26	−0.09	1.48	−1.48
27	0.44	0.74	−0.74
28	−1.46	0.91	−0.91
29	0.71	0.93	−0.93
30	1.50	−1.10	1.10
Mean	0.06	1.00	−1.00
SD	1.12	1.00	1.00

when estimating item and person parameters across the conditions simulated by Sass et al. The greatest difficulty for each of these methods was in recovering θ estimates for individuals in the tails of the distribution (e.g., above 2).

Methods for Estimating IRT Model Parameters in the Presence of a Nonnormal Latent Trait

Given the clear evidence that estimation problems can arise when the latent trait is not normally distributed, researchers have developed a number of methods that might be more appropriate in such conditions. Of particular interest in the current study are Ramsay curve item response theory (RC-IRT) and nonparametric Bayesian estimation (NBE). RC-IRT has been fairly well studied in recent years (Woods, 2006, 2007, 2008; Woods & Linn, 2009; Woods & Thissen, 2006) and has been found to provide accurate IRT model parameter estimates under a variety of conditions (e.g., with a skewed latent trait distribution, 30 or more items). More recently, the NBE method was introduced as an alternative for use with Rasch models when assumptions about the distribution of the model parameters cannot be made (San Martin, Jaran, Rolin, & Mouchart, 2011). In contrast to RC-IRT, however, NBE has not been thoroughly examined under a variety of conditions. San Martin et al. (2011) did conduct a small simulation study using both a normally distributed and a bimodal distribution of θ, and they found that NBE yields accurate person and item parameter estimates. However, as noted by a number of researchers (e.g., de Ayala, 2009; Kirisci & Hsu, 1995; Stone, 1992), standard approaches for estimating IRT parameters work reasonably well when the distribution is symmetric, which would include the bimodal distribution simulated by San Martin et al. Therefore, relatively little is known about how well NBE might recover the model parameters when θ or the set of item parameters are skewed. Following is a description of the NBE method used here, given that it is the newest and least studied. Details of the MML, MAP, RC-IRT, and Bayesian Markov chain Monte Carlo (MCMC) approaches to estimation, which are also included in this study, are not provided here as they have been thoroughly discussed elsewhere and are very commonly used in the IRT literature (de Ayala, 2009; Fox, 2010; Woods, 2015).

Nonparametric Bayesian Estimation

Another alternative method for IRT model parameter estimation in the context of nonnormal latent traits that has been suggested in the literature is NBE, which was described by San Martin et al. (2011). These authors focused on the Rasch model, linking the item responses under this model to the Bernoulli distribution as follows:

$$\left(x_{ij}|\theta_i b_j\right) \sim \text{Bernoulli}\left(\frac{e^{\left(\theta_i - b_j\right)}}{1+e^{\left(\theta_i - b_j\right)}}\right), \tag{2}$$

where $x_{ij} = 1$ indicates that subject i correctly answers item j, and $x_{ij} = 0$ indicates an incorrect response; θ_i is the level of the latent trait for subject i; b_j is the difficulty for item j.

The entire item response pattern for subject i, Y_i, is characterized by the distribution

$$P(Y_i = y_i|b_{1:J}, G) = \int \left\{ \prod_{1 \le i \le n} P\left[Y_{ij} = y_{ij}|\theta_i, b_j\right] \right\} dG(\theta), \tag{3}$$

where Y_i is the response pattern for subject i; $b_{1:J}$ is the set of difficulty values for the J items; and G is the probability distribution of mixtures of the latent trait, θ.

San Martin et al. (2011) noted that when θ cannot be assumed to follow the normal distribution, estimates of item and person parameters are likely to be compromised in the form of bias, which is an issue that has already been described in some detail above. Their solution to this problem was to apply the Bayesian estimation paradigm to the problem, but to do so in a way that imposes the least restrictive prior information on the model parameters possible. Briefly, in practice Bayesian estimation involves the combination of information from the data (the likelihood) with prior information about the model parameters, in order to obtain a posterior distribution for each parameter. The point estimates for the parameters are drawn from this posterior distribution in the form of the mean, median, or mode. For an extensive discussion of general Bayesian IRT modeling, the interested reader is referred to Fox (2010).

When the Bayesian methodology is applied to the Rasch model, prior distributions must be given for both θ and b_j. Any reasonable prior can be provided for these parameters based on information that the researcher has about the parameters and their distributions. Most often in practice the normal distribution with mean of 0 and variance of 1 is used as the prior for both θ and b_j (Fox, 2010). As noted previously, however, IRT model parameters may not always follow the normal distribution, calling into question the use of normal priors in combination with a normal based likelihood function, which is standard practice. The NBE approach described by San Martin et al. (2011) can be based on either the Dirichlet process (DP), which is a prior probability distribution on a space of probability distributions, or Polya trees (PT), which involves the probabilistic partitioning of observations into increasingly homogeneous bins based on their distributional form (Lavine, 1992).

The DP is characterized by the parameters (M, G_0), where $M \in \mathbb{R}$, and $G_0 \in P(\mathbb{R})$. In other words, the probability distribution, G, generating the individual latent trait θ, is assumed to have come from some set of distributions contained within the DP with the uncertainty regarding the exact distribution being reflected in the prior distribution on $P(\mathbb{R})$. The central tendency of this distribution is contained in the prior distribution of M. So, for example, the researcher could specify as a prior distribution for θ the DP process based on a mixture of normal distributions, rather than a single normal distribution for the entire set of data. This mixture of normal distributions should serve to better characterize a parameter that does not in fact conform to a pure normal than would normal based methods such as MML, or even a standard Bayesian approach that uses a single normal prior (San Martin et al., 2011). It is also important to note that DP is not limited to the normal distribution, but rather can accommodate any distributional form that the researcher believes to be appropriate. Finally, NBE relies on the MCMC algorithm, which is very commonly used to obtain posterior distributions for model parameters in Bayesian statistics (Geyer, 1992).

The PT priors come from a sequence of m partitions in which nested subsets of the data are obtained. The probability distribution G for the latent trait θ is then distributed as PT (A, Π), where $A=\{\alpha_\varepsilon\}$ and Π contains the set of partitions making up the PT. The term α_ε controls the level of refinement desired in the PT solution (i.e., the number of splits to be done), and is itself given a prior distribution. In many instances it may be difficult for a researcher to specify a single distribution around which the PT will be centered. In such cases, a mixture of PT (MPT) can be used by setting a hyperprior on the centering parameter (i.e., mean) of the PT (Hanson & Johnson, 2002). As an example, if for a standard PT the prior distribution for the mean of G is N(0, 100), then for an MPT the prior distribution for G would become N(η, 100), where η is itself a prior distribution, such as N(0, 500). Through the use of MPT, the resulting posterior distribution will not be tied to a single centering prior distribution, and should therefore accommodate a wide array of potential distributions to be found in a population (San Martin et al., 2011).

San Martin et al. (2011) conducted a small simulation study in order to assess the performance of the NBE approach in terms of estimating both θ and b_j for the Rasch model. The latent trait used in their study was simulated from a mixture of two normal distributions, which took the form $0.5N(-1, 0.5^2) + 0.5N(2, 0.25^2)$ and yielded a bimodal distribution. For each replication, 250 examinees were simulated, and either 2, 4, 10, or 40 items were used. The outcome of interest was the Kolmogorov distance between the posterior mean of the Bayesian process estimating θ and the data generating distribution itself. The priors used in this simulation were centered at the $N(\mu, \sigma^2)$, where the prior on μ was $\mathrm{N}(0, 100)$, the prior on σ^{-2} was $\Gamma\left(\frac{2.01}{2}, \frac{0.01}{2}\right)$, and the prior on α_ε was $\Gamma(2.0, 0.2)$. In order for the model to be identified, the difficulty for item 1 was set equal to 0, and the prior distribution of the remaining item parameters was $\mathrm{N}\left(0, 10^3 x I_{n-1}\right)$. The results of this study demonstrated that NBE worked best for cases with larger numbers of items, and that the posterior distribution was sensitive to the prior distribution used. This latter result may be due in part to the relatively small sample size used in the simulation. In addition, the MPT approach outperformed the other methods for NBE with the Rasch model.

Study Goals

The current study was designed to extend work in the area of IRT model parameter estimation in the presence of nonnormal latent trait and item difficulty patterns in several ways. First, the NBE technique was compared with other previously studied methods, under a much broader array of conditions than it has been examined with heretofore. Second, the performance of RC-IRT (as well as the other methods studied here) was considered in light of interactions between the item and θ distributions. Prior work on this proven method has focused primarily on its performance with various distributions of the latent trait, but not with regard to different item difficulty distributions. Third, the standard Bayesian estimation method (MCMC) was included in the study as well. Other approaches use Bayesian methods to estimate the latent trait (i.e., MAP and EAP), but little previous research has examined performance of the MCMC method for estimating both person and item parameters when item difficulty and/or θ are not normally distributed. Fourth, this study examined the performance of the various methods with respect to both item and latent trait recovery. Much of the prior work, particularly with NBE, has focused on the latent trait but not item recovery. Finally, the current study broadens the number of items, sample size conditions, and distribution of the latent trait as compared to previous research, in an attempt to identify where these methods perform optimally, and where they have problems.

Based on previous research, it was possible to develop hypotheses regarding what might be expected in this study. First, it was anticipated that NBE would outperform the MCMC approach when the model parameters were skewed, as the former does not rely so heavily on the form of the prior distribution of these parameters as does the latter. Our second hypothesis was that RC-IRT would perform better than MML or MCMC for item parameter estimation, and better than MML, MAP, or MCMC for θ estimation because it does not rely on a normality assumption, nor on normal priors. Given its relatively recent development and the need for more extensive examination of its characteristics, we are not able to develop any hypotheses regarding the performance of NBE vis-à-vis RC-IRT, other than to surmise that both methods would perform relatively better than the other methods studied here. Neither technique has been examined under several of the conditions included here, so it was not clear how accurate their estimates would be when compared with one another in these cases.

Method

In order to address the goals of this study, a Monte Carlo simulation methodology was used. Several factors were manipulated (as described below), and completely crossed with one another. For each combination of conditions, 1,000 replications were simulated. Data generation and analysis were conducted using the R software system (R Core Development Team, 2014) and RCLOG (Woods, 2006). Following are descriptions of how item difficulty and person latent trait parameters were generated, how model parameters were estimated, and how the quality of these estimates was ascertained. The manipulated factors in this study were item difficulty parameter distribution (3 levels) by latent trait distribution (5 levels) by method of estimation (5 levels) by sample size (4 levels) by number of items (3 levels), for a total of 900 separate conditions.

Item Difficulty Parameter

The data generation for the current study was based on work by Sass et al. (2008). All data were generated using the Rasch model, with item difficulty parameters drawn from Sass et al. and item discrimination values set to 1. The item difficulty values were generated under three different conditions: normally distributed, positively skewed, and negatively skewed. The values under each condition appear in Table 1. Under the normally distributed condition, item difficulty values were drawn from the standard normal distribution, and designed to give maximal information for examinees with latent trait values in the middle of the distribution. The positively skewed item difficulty values were drawn from the gamma distribution with a mean of 1 and a standard deviation of 1, and provided maximal information about examinees with latent traits near -1. Items from the negatively skewed difficulty distribution had difficulty parameter values that were identical to those in the positively skewed case, but with the signs reversed. Three conditions for number of items were simulated, including 10, 20, and 30. This condition was manipulated in the current study, as it has been shown to be a salient factor in the performance of RC-IRT (Woods, 2015) and NBE (San Martin et al., 2011). For the 10 items condition, the difficulty parameters for the first 10 items in Table 1 were used, whereas for the 20 items condition, the difficulty values for the first 20 items were used in the simulations, and for the 30 items condition all of the item difficulty parameter values were used to generate the data.

Latent Trait

The latent trait was generated to be from one of five distributions: Standard Normal, bimodal mixture of normals ($0.5N(-1, 0.5^2) + 0.5N(2, 0.25^2)$), skew of 1.5, 2.5, or 3.5. The bimodal distribution was used previously in San Martin et al. (2011) with the NBE, and skewed distributions have been studied with RC-IRT and the MML/MAP estimation methods previously. However, these previous studies did not utilized latent traits that were skewed as severely as 3.5. Thus, the current study was designed to build on prior work by using similar conditions (e.g., bimodal, skew of 1.5) to prior studies, but also to apply new distributional conditions (e.g., skew of 3.5) and the older conditions to new methods (e.g., bimodal with RC-IRT, skewed distributions with NBE). Samples of 250, 500, 1,000, and 2,000 examinees were simulated.

Estimation Methods

Person latent trait parameters were estimated using MML, MAP, RC-IRT, NBE, and Bayesian IRT (MCMC; Fox, 2010). Item difficulty parameter estimation was carried out using MML, RC-IRT, NBE, and MCMC. MAP was used with a standard normal prior, and model estimation was carried out using the `RM` and `PP_4pl` functions of the `eRm` and `PP` R libraries. MCMC was carried out using standard normal priors for both latent trait and item difficulty parameter values, as is common in the literature (e.g., Fox, 2010; Kim & Bolt, 2007). Two chains of 20,000 iterations each were generated using MCMC, with a burn in period of 1,000 iterations, and thinning set to 20. Prior to conducting the simulation portion of the study, several test analyses were conducted using data generated under the conditions described above so that parameter estimation convergence could be determined based on trace plots and autocorrelation functions. It was found that using the MCMC settings described above, model convergence was attained within 1,000 iterations in all cases. Therefore, we were very confident that using 20,000 iterations was sufficient for appropriate parameter estimation under all conditions studied here. The MCMC modeling was done using the `MCMCirt1d` function in the `MCMCpack` R library.

The NBE estimation was done using the `DPMrasch` function in the `DPpackage` R library. Based on prior results demonstrating its strong performance (San Martin et al., 2011) the MPT method was used for NBE, with the nonparametric priors centered at the $N(\mu, \sigma^2)$ distribution. For μ, the $N(0, 100)$ prior distribution was used, whereas the prior for σ^{-2} was distributed as $\Gamma\left(\frac{2.01}{2}, \frac{0.01}{2}\right)$, in keeping with San Martin et al. (2011). The MCMC algorithm was used to obtain parameter estimates, using 20,000 iterations with a burn-in of 1,000, and thinning of 20. As with the MCMC approach, these settings were tested on data simulated using the methods described above. It was found that convergence was always obtained by 1,000 iterations. Therefore, it was determined that using 20,000 iterations with a burn-in of 1,000 would ensure model parameter estimation convergence.

The use of RC-IRT to fit the models was based on the simulation work of Woods (2006), and subsequent recommendations for practice (Woods, 2015). Specifically, 25 different models were fit to each simulated dataset using the RCLOG software, with 5 number of knots used (2, 3, 4, 5, and 6), crossed with the 5 separate orders of the polynomials (2, 3, 4, 5, and 6). In addition, 121 quadrature points between -6 and 6 were used in the estimation process. A prior standard deviation of 75 was used with the normal prior applied to the spline parameters. The Hannan–Quinn (HQ) criteria and the Kolmogorov–Smirnov test were used to identify the optimal solution for each replication. Specifically, following the recommendations of Woods, the model with the smallest HQ statistic was selected as the provisional best fit to the data. The resulting latent trait from this model was then compared to the normal distribution using the Kolmogorov–Smirnov test ($\alpha = 0.05$). If the result was statistically significant, the result was retained, otherwise the normally distributed latent trait estimates were used. The item and person parameter estimates for this optimal solution were then used as the RC-IRT estimates for purposes of calculating the outcome variables for the simulation study (described below), and for comparison with the other estimation methods.

Study Outcomes

The quality of the item and person parameter estimates were assessed and compared in two ways. First, the absolute value of the estimation bias, sometimes referred to as mean absolute error (*MAE*), was calculated for each item and person estimate. For the item difficulty values, this calculation was

$$MAE = \left|b_j - \hat{b}_j\right|, \tag{4}$$

where b_j is the population value of difficulty parameter for item j; $\hat{b}_j$ is the estimate of difficulty parameter for item j.

Similarly, the absolute bias in the latent trait estimates was calculated as

$$MAE = \left|\theta_i - \hat{\theta}_i\right|, \tag{5}$$

where θ_j is the population value of latent trait parameter for person i; $\hat{\theta}_j$ is the estimate of latent trait parameter for person i.

The second outcome of interest in the current study was the mean squared error (*MSE*) of the model parameter estimates, which for item estimates took the form:

$$MSE_j = \frac{\sum_{r=1}^{R}\left(b_j - \hat{b}_j\right)^2}{R}. \tag{6}$$

Terms are as defined above, with the addition that R is the number of replications. For latent trait estimates, *MSE* is calculated as

$$MSE_j = \frac{\sum_{r=1}^{R}\left(\theta_i - \hat{\theta}_i\right)^2}{R}. \tag{7}$$

In order to determine which of the manipulated study factors were significantly related to the outcomes, a repeated measures analysis of variance (ANOVA) was used. The dependent variable was absolute bias (analyses were run separately for item and person latent traits), the repeated measures factor was estimation method, and the between replication factors were distribution of the item difficulty parameters, difficulty of the latent trait, number of examinees, and number of items. With regard to the item parameter estimates, a separate ANOVA was run for each item and the results were compared with one another in terms of the manipulated study effects that were identified as being related to the outcome variable, MAE.

Results

The results of the simulation study for item difficulty and person parameter estimation are presented in the following section of the article. Difficulty parameter estimation results were obtained for all items in each replication, and as noted above ANOVA results were obtained for each item. However, given the many combinations of study conditions, it was not feasible to present results for each item. Furthermore, in terms of the factors that were identified as being significantly associated with MAE, the ANOVA results were very similar across the items. Therefore, in order to present study results fully reflective of item difficulty parameter values for the various study conditions, difficulty estimation outcomes for Items 2 and 10 were selected for presentation below. These items were selected because they took a range of population values across the item difficulty parameter distributions used. Thus, by focusing on these two items, we had combinations of items with difficulty values near 0, as well as both extremely positive and extremely negative for the normal, negative skewness, and positive skewness item difficulty distributions.

Item Difficulty Estimation

The ANOVA used to identify significant manipulated study factors with regard to item parameter MAE found that estimation method by item difficulty distribution by latent trait distribution ($F_{16,248} = 5.555, p < .00001, \eta^2 = 0.264$), and the interaction of estimation method by sample size ($F_{3,124} = 51.982, p < .00001, \eta^2 = 0.557$) were statistically significant for the target items. Note that the results for both of the target items were very similar to one another, so only the ANOVA results for Item 2 are presented above. Figure 1 contains bar charts for MAE by estimation method, distribution of the latent trait, and distribution of item difficulty parameters, for Items 2 (top panel) and 10 (bottom panel). With respect to Item 2, when the distribution of θ was normal, MML, NBE, and RC-IRT had very similar MAE values across the distribution of the item difficulty conditions. Furthermore, these were lower than those of MCMC when the item difficulty distribution was either normal (b = 1.08), or negatively skewed (b = −0.49). However, when the item distribution was positively skewed (b = −0.49), the MCMC bias was comparable to that of the other methods. When the data were bimodal, the lowest MAE results were associated with NBE and MML for both the normal (b = 1.08) and positively skewed item difficulty conditions (b = 0.49). Item difficulty estimates from RC-IRT exhibited the largest MAE in these conditions. On the other hand, when the item difficulties were negatively skewed in the population (b = −0.49) and the latent trait was bimodal, the four estimation methods yielded very similar levels of MAE for Item 2. Finally, when the latent trait was skewed, the pattern of difficulty parameter MAE for Item 2 was very similar to that in evidence when the latent trait was normally distributed.

MAE results for Item 10 appear in the bottom panel of Figure 1. When θ was normally distributed and the item parameters were also normally distributed (b = −0.24), the MAE present in estimates produced by MML, RC-IRT, and NBE was generally comparable, with slightly higher values for MCMC. However, when the item parameters were either negatively (b = 2.30) or positively (b = −2.30) skewed, MAE for MCMC was much larger than that of the other methods, and RC-IRT exhibited greater MAE than did either MML or NBE, both of which yielded results similar to those in the normal item difficulty distribution condition. A similar pattern of MAE was evident when the latent trait followed a bimodal distribution, except that the value for RC-IRT was slightly higher than the others in the normal difficulty item condition (b = −0.24), and was nearly comparable to that of MCMC in the negative item skewness case (b = 2.30). When the latent trait was skewed, the pattern of item difficulty MAE was similar to that when θ was normally distributed.

Figure 2 contains the MSE for the item difficulty estimates by estimation method, distribution of item difficulty parameters, and distribution of the latent trait. Panel 1 of Figure 2 includes the MSE for item difficulty parameter estimates of Item 2 for

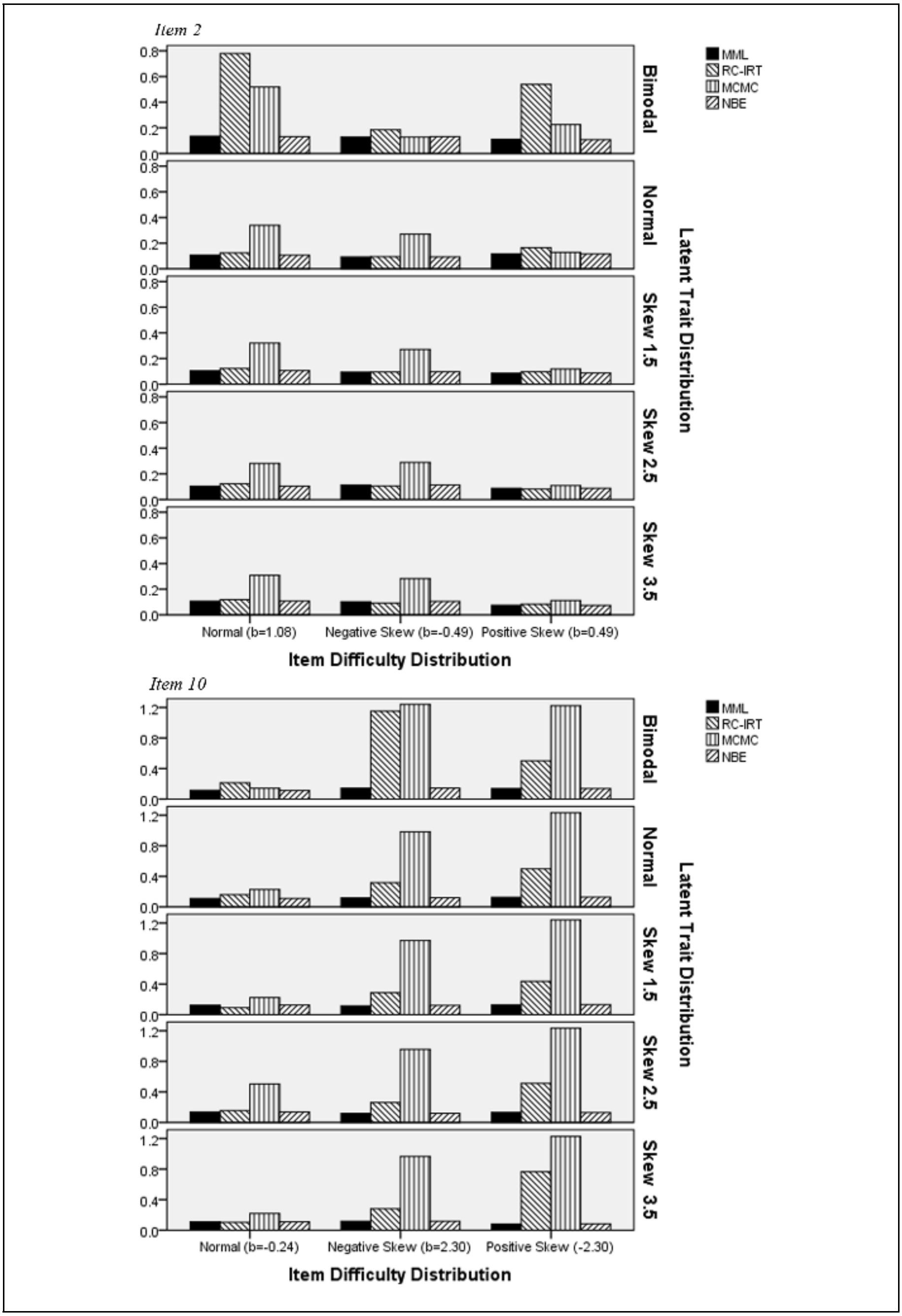

Figure 1. MAE of difficulty estimates for Items 2 and 10 by estimation method, latent trait distribution, and item difficulty distribution.

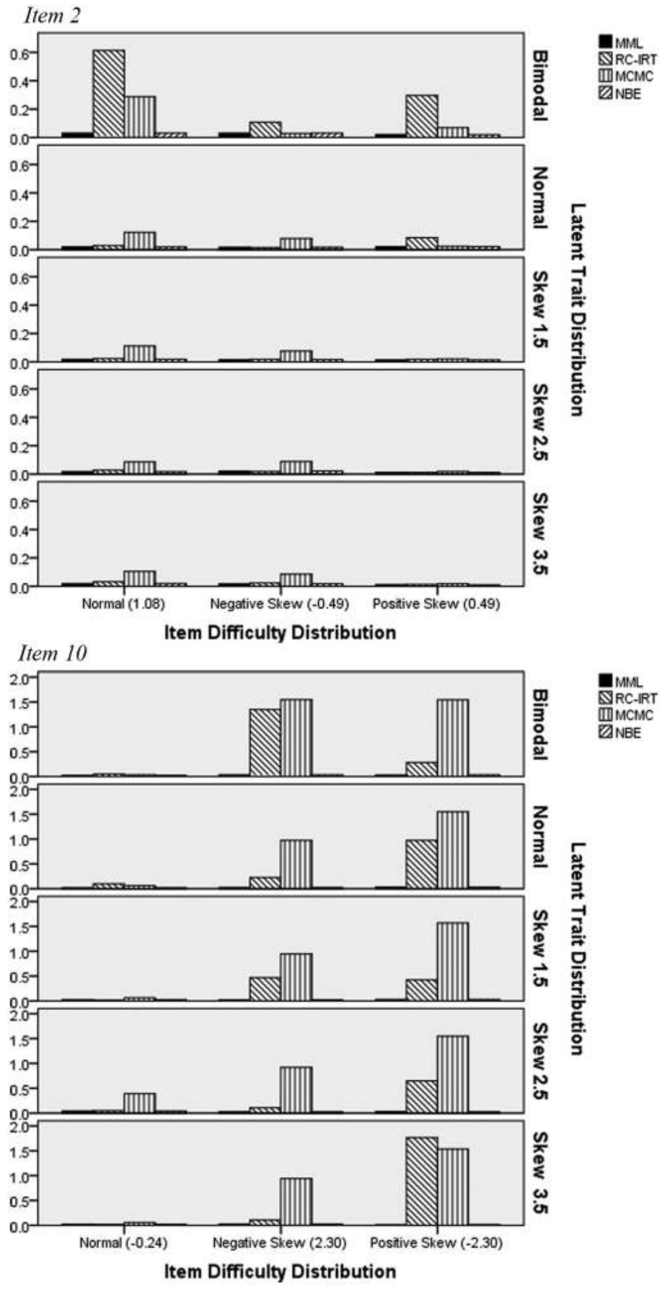

Figure 2. MSE of difficulty estimates for Items 2 and 10 by estimation method, latent trait distribution, and item difficulty distribution.

Table 2. MAE and MSE for Difficulty Estimates of Items 2 and 10 by Estimation Method and Sample Size (*N*).

	MLE		RC-IRT		MCMC		NBE	
N	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Item 2								
250	0.16	0.04	0.25	0.12	0.26	0.09	0.16	0.04
500	0.11	0.02	0.21	0.10	0.22	0.07	0.11	0.02
1,000	0.08	0.01	0.15	0.07	0.14	0.08	0.08	0.01
2,000	0.06	0.01	0.14	0.06	0.14	0.06	0.06	0.01
Item 10								
250	0.19	0.06	0.59	1.10	0.66	1.16	0.19	0.06
500	0.14	0.03	0.36	0.29	0.55	0.42	0.14	0.03
1,000	0.09	0.01	0.31	0.20	0.48	0.30	0.09	0.01
2,000	0.06	0.01	0.27	0.16	0.45	0.26	0.06	0.01

each method, by the distributions of θ and item difficulty. The pattern of results for MSE was similar to those of MAE across the simulated conditions. In sum, when the data were simulated to be bimodal, RC-IRT had the largest MSE values among the methods for Item 2, with the greatest deviations for the normal item distribution case (b = 1.08). In contrast, when the data were simulated to be skewed to some degree, MSE was greatest for MCMC, particularly in the normal (b = 1.08) and negatively skewed (b = −0.49) conditions. Finally, across conditions, MSE was lowest for the skewed conditions when the item parameter distribution was positively skewed (b = 0.49). It should be noted that in this case, the skewness of the latent trait and the item difficulty parameters was in the same direction, that is, positive.

The bottom panel of Figure 2 contains the MSE for Item 10 by method, latent trait skewness, and item difficulty skewness. As was the case for Item 2, the pattern of results for MSE with Item 10 was very similar to the pattern seen in the bottom panel of Figure 1 for MAE. Both MCMC and RC-IRT exhibited the greatest MSE when the item difficulty parameters were positively skewed (b = −2.30). In addition, for most of the simulated conditions MSE was largest for MCMC, followed by that of RC-IRT. The lone exception to this pattern occurred when θ was most positively skewed, in which case RC-IRT had the larger MSE value. When the latent trait was normally distributed, all of the methods had comparable MSE values, except for MCMC for which it was larger when skewness = 2.5. Both MLE and NBE exhibited the lowest, and very comparable, MSE values across all conditions for Item 10. Indeed, particularly relative to the other methods, the performance of neither MLE nor NBE appeared to be influenced by the simulated conditions.

The MAE and MSE values associated with each method by sample size for Items 2 and 10 appear in Table 2. From these results, it appears that across other simulated conditions MAE and MSE both declined with increasing sample size for all the methods, for both items. This rate of decline was comparable for MLE and NBE, and

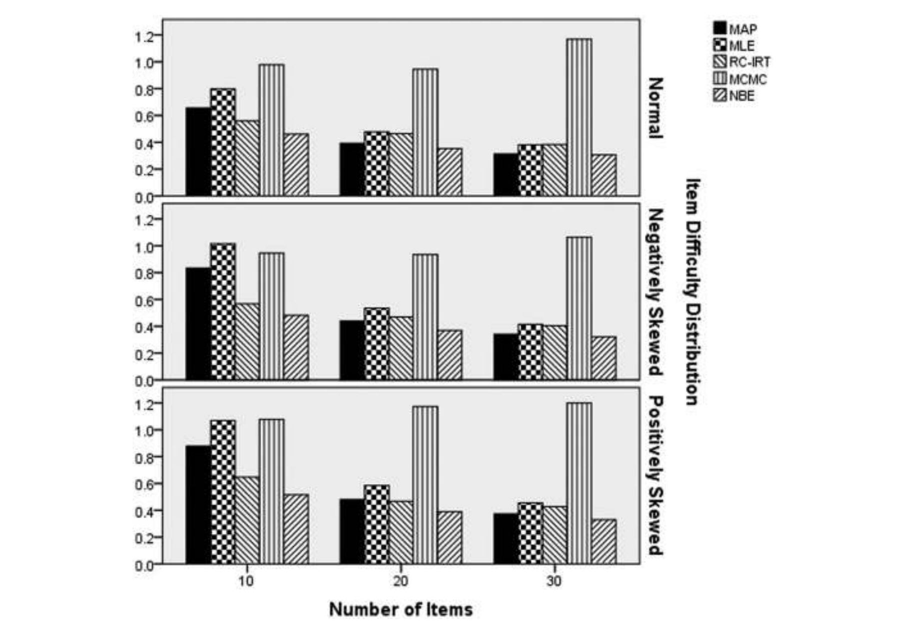

Figure 3. MAE of latent trait estimate by estimation method, item difficulty parameter distribution, and number of items.

somewhat greater for both RC-IRT and MCMC for both items. In other words, sample size was somewhat more salient for the performance of RC-IRT and MCMC than for either MLE or NBE.

Latent Trait Estimation

In order to identify the important main effects and interactions of the manipulated conditions with respect to the MAE for the latent trait estimates, a single repeated measures ANOVA was used, in which the within subjects factor was estimation method, and the between subjects factors were latent trait distribution, item difficulty distribution, number of items, and sample size. By including all of the study conditions in a single ANOVA, we were able to identify those that were significantly related to estimation MAE, and thus more broadly understand the factors that affected this outcome variable. The ANOVA identified the interaction of method by item difficulty distribution by number of items ($F_{16,496} = 8.479, p < .00001, \eta^2 = 0.215$), and the interaction of method by latent trait distribution ($F_{32,496} = 6.165$, $p < .00001, \eta^2 = 0.285$) as significantly related to MAE. All other terms in the model were either not statistically significant, or were subsumed in one of these interactions. Figure 3 contains MAE for estimates of θ for each method by the distribution of item

Table 3. MAE for Latent Trait Estimates by Method, Item Difficulty Distribution, and Latent Trait Distribution.

Latent trait distribution	Item difficulty distribution	MAP	MML	RC-IRT	MCMC	NBE
Normal	N	0.43	0.41	0.42	0.42	0.40
	NS	0.53	0.65	0.45	0.88	0.43
	PS	0.54	0.66	0.52	0.98	0.43
Bimodal	N	0.52	0.63	0.71	1.61	0.40
	NS	0.56	0.68	0.75	1.55	0.42
	PS	0.75	0.92	0.76	1.66	0.45
Skew 1.5	N	0.42	0.51	0.40	0.52	0.39
	NS	0.53	0.64	0.41	0.91	0.41
	PS	0.53	0.70	0.43	1.15	0.41
Skew 2.5	N	0.51	0.62	0.39	0.80	0.40
	NS	0.64	0.75	0.41	1.08	0.40
	PS	0.64	0.79	0.44	1.38	0.43
Skew 3.5	N	0.75	0.81	0.41	1.34	0.40
	NS	0.94	1.05	0.40	1.56	0.41
	PS	1.03	1.12	0.46	1.77	0.42

difficulty parameters and the number of items. For all methods, MAE declined as the number of items increased, except for MCMC, on which the number of items did not appear to have an impact. The number of items had a particularly strong impact on MAP and MLE estimation. For 10 items, these two estimation approaches exhibited much greater MAE than either RC-IRT or NBE. However, for 30 items MAP and MLE yielded comparable MAE to RC-IRT across the distribution of the item difficulty parameters. Latent trait MAE results were somewhat lower for NBE than for the other approaches, regardless of the item difficulty distribution. In addition, MAP consistently yielded lower MAE than did MLE. With respect to the impact of item difficulty distribution on the estimation of θ, MAE values were lower in the normal condition for MAP and MLE, particularly for 10 items. For larger number of items, the difference in MAE for these two approaches across item difficulty distributions declined. On the other hand, for NBE and RC-IRT there were no differences in MAE for θ based on item difficulty distribution. MCMC yielded somewhat lower MAE values for negatively skewed item difficulty parameters than for the other two conditions.

Table 3 contains MAE values for the latent trait by estimation method, distribution of item difficulty parameters, and the distribution of the latent trait. Across all simulated conditions MCMC yielded the most biased estimates, except when both the latent trait and the item difficulty parameters were simulated to be normally distributed, in which case all of the methods performed similarly with respect to MAE. When the latent trait was normally distributed and the item difficulty parameters were not, MAE for all methods increased compared to the normally distributed

Table 4. MSE for Latent Trait Estimates by Method, Item Difficulty Distribution, and Latent Trait Distribution.

Latent trait distribution	Item difficulty distribution	MAP	MML	RC-IRT	MCMC	NBE
Normal	N	0.42	0.51	0.48	1.58	0.46
	NS	0.89	1.09	0.50	1.81	0.50
	PS	0.92	1.12	0.53	1.94	0.50
Bimodal	N	0.69	0.84	0.73	4.01	0.48
	NS	0.97	1.17	0.79	3.73	0.50
	PS	1.48	1.80	0.88	4.17	0.56
Skew 1.5	N	0.49	0.53	0.48	2.43	0.45
	NS	0.89	1.09	0.51	2.70	0.46
	PS	0.90	1.07	0.52	3.12	0.51
Skew 2.5	N	0.84	1.03	0.49	2.75	0.46
	NS	1.01	1.21	0.53	3.43	0.46
	PS	1.10	1.32	0.55	3.99	0.50
Skew 3.5	N	1.17	1.34	0.49	3.77	0.45
	NS	1.37	1.65	0.53	4.12	0.47
	PS	1.41	1.68	0.56	4.47	0.52

difficulty values, with the smallest increase occurring for NBE. The impact of negatively versus positively skewed item difficulty parameters was similar when the latent trait was normally distributed. When the data were generated from the bimodal distribution, all of the methods exhibited higher latent trait MAE when compared to the normally distributed latent trait, except for NBE, which performed similarly in the two conditions. Again, MCMC produced the most biased θ estimates, followed by those of RC-IRT. The relationship between the skewness of θ and the value of MAE was positive for most of the estimation methods such that greater bias was found with a more skewed distribution. The exception to this pattern occurred for NBE, and for RC-IRT when the items were normally distributed. However, when the items were simulated from a skewed distribution, MAE for RC-IRT was much greater than in the normal case. In addition, for all of the methods, MAE for θ estimates was highest when both the latent trait and item difficulty parameters were positively skewed.

Table 4 includes MSE values of the latent trait estimates by estimation method, and the latent trait and item distributions. MSE for all methods was lowest when both the latent trait and the item difficulty parameters were normally distributed. Indeed, in this case MAP yielded the lowest MSE indicating that its estimates were closest to the data generating values. When the item parameters were skewed, MSE for all of the methods increased, with the smallest such increases occurring for NBE, followed by RC-IRT. When the latent trait took a bimodal distribution, NBE consistently had the lowest MSE, followed by RC-IRT. MSE in the θ estimates increased concomitantly with increases in the skewness of the latent trait for all of the methods except NBE and

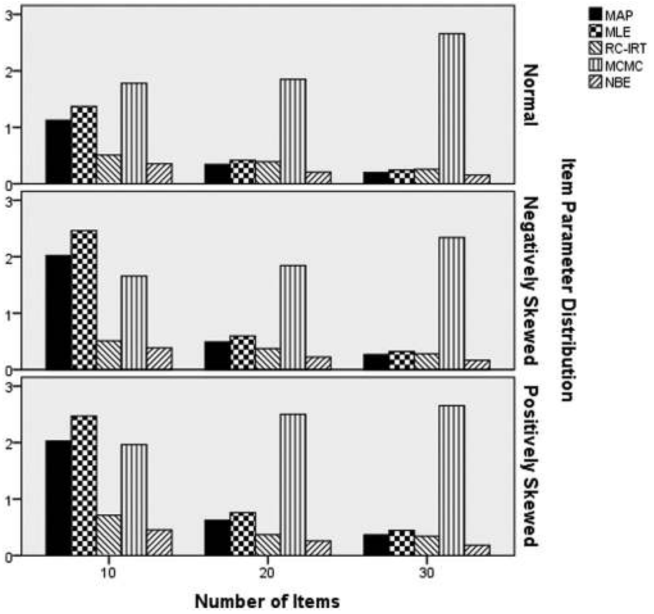

Figure 4. MSE of latent trait estimate by estimation method, item difficulty parameter distribution, and number of items.

RC-IRT, both of which had MSE values similar to those they exhibited in the normal distribution condition, regardless of the degree of bias. Finally, all of the methods studied here, including RC-IRT and NBE, had somewhat larger MSE values when both the latent trait and the item parameters were positively skewed (see Figure 4).

Discussion

The purpose of this study was to examine the performance of several methods of item and person parameter estimation for the Rasch model when the latent trait and item difficulty parameters were not normally distributed in the population. Two goals were central to this work: (1) investigation of a relatively new method for Rasch model parameter estimation based on the nonparametric Bayesian paradigm and (2) furthering prior work with RC-IRT by including nonnormal item difficulty parameters as well as a nonnormal latent trait. The results presented above demonstrated that when the latent trait and item difficulty parameters were normally distributed, all of the methods provided similarly accurate latent trait estimates, which has been demonstrated to be the case in prior studies as well. However, again similar to prior research results, when the latent trait was not normally distributed normal based methods such as MML, MAP, and MCMC with normal priors exhibited more bias and higher MSE values than either RC-IRT or NBE. In turn, NBE was consistently the best performer in terms of parameter estimation accuracy as measured by MAE and MSE. Even for the bimodal distribution, which created problems for RC-IRT (which performed well with a skewed latent trait), NBE was able to yield accurate person parameter estimates. In terms of item parameter estimation, NBE again was always among the most accurate, if not the most accurate method studied here, regardless of the simulation condition. In addition, item difficulty estimation for the other methods was generally more accurate when the item parameters were normally distributed as a whole, and when the specific item parameter was less extreme (i.e., closer to 0). This last result was particularly in evidence for MCMC, which had the greatest difficulty estimating more extreme item parameter values across conditions.

Implications for Practice

The results of this study present several implications for researchers and measurement practitioners alike. First, the NBE method shows great promise for use in estimating both item and person parameters in the context of Rasch modeling when either or both are not normally distributed. These results are in keeping with those in San Martin et al. (2011), but extend on that earlier work both in terms of the types of nonnormal distributions that were used, and the inclusion of nonnormal item difficulty parameters, per Sass et al. (2008). It must be noted, however, that despite its strong performance in the context of the Rasch model used in this study, the NBE approach is limited in practice because it is currently only available for fitting the Rasch model with either dichotomous or Poisson data. San Martin et al. (2011) discuss the fact that work is underway to extend this estimation technique to more complex IRT models, but currently such are not available. Thus, when considering alternatives to NBE that may be useful for nonnormal data conditions and more complex (e.g., 2-parameter logistic or 3-parameter logistic) models, the results presented here provide further evidence that the RC-IRT approach might be an attractive alternative when the latent trait is skewed. In those cases, RC-IRT was nearly as accurate in terms of both item and person parameter estimation as NBE under most conditions simulated here. However, when the latent trait was bimodal, RC-IRT exhibited some difficulty when estimating both item difficulty and the latent trait of interest. Thus, for more complex models than the Rasch, this study suggests that MAP and MML may be preferable to use with a bimodal distribution. In making this recommendation, we are very careful to note that the current study was limited to the Rasch model. Thus, it is possible that when estimating person and item parameters in the nonnormal conditions simulated here, performance of the various methods will not follow the patterns seen here when the underlying model is more complex. It is for this reason that we say the current study only provides preliminary evidence for more complex models. Clearly, more research needs to be done investigating the performance of these estimation methods with more complex models. However, because one of our primary goals was to study NBE in a wider range of conditions than has been done heretofore, we were limited to using the Rasch model. Nonetheless, we believe that the current results are not only useful for those interested in using the Rasch model with nonnormal person and item parameters, but also as an extension of prior work with RC-IRT by exploring its performance with nonnormal item difficulty parameters in conjunction with a nonnormal latent trait.

Limitations and Directions for Future Research

While the current study extends prior work in the area of IRT model parameter estimation in the context of nonnormal latent trait and item difficulty distributions, there are a number of areas in which it needs to be further developed in future work. First, the current study is limited to the Rasch model, as noted above. While NBE is only available for Rasch estimation as of this writing, plans exist to extend it to the 2PL and 3PL models, so that future work should investigate its performance for these. This study demonstrates its great promise under a variety of conditions with the Rasch model, but it is not clear whether this will be the case for more complex IRT models. In addition, the current work demonstrated the difficulty that MCMC estimation has in most instances when the latent trait and/or item difficulty parameters are not normally distributed, and a normal prior is used. However, it is not clear how MCMC estimation might work were more accurate priors placed on the model parameters. It has been shown that another Bayes based method for estimating the latent trait, EAP, performed optimally when the nonnormal distribution used to generate the data was also used as the prior for θ (Sass et al., 2008). While this may not be particularly realistic in practice, it is possible that a researcher could examine the distribution of observed scores from a scale, and use these as a starting point to determine the priors for MCMC estimation. Similarly, an examination of the proportion of correct responses to a dichotomous item could be used for a similar purpose in ascertaining priors for item parameters. In this way, the researcher might be able to develop more accurate priors for use with MCMC, than the normal based distributions used in this study. Thus, future work with MCMC and nonnormal data could make use of more accurate such priors. Given its promise, future research should also incorporate DC-IRT (Woods & Linn, 2009) estimation with nonnormal latent trait and item difficulty parameters. As noted above, this methodology is somewhat limited in terms of practical utility by the lack of an easy to use software interface. However, as such software becomes available researchers will need to know how well DC-IRT performs under conditions similar to those simulated here. Early work with this method by Woods and Linn suggests that it is a promising approach. Finally, Sass et al. (2008) examined the recovery of latent trait estimates for individuals with extreme values of θ. Given the dual focus on recovery of both person and item parameters for several estimation methods, it was felt that including estimation recovery results for examinees with such extreme estimates was beyond the scope of the current work. However, future research should investigate the ability of each of these methods, particularly RC-IRT and NBE, to accurately estimate extreme values of the latent trait under the various conditions simulated here.

Conclusions

The current study shows that under a variety of skewed distributions, as well as with bimodal data, NBE may be the most accurate procedure available for Rasch model parameter estimation. It proved to be most accurate in virtually all scenarios that were simulated here, and is not difficult to use with the R library DPpackage. Thus, researchers interested in fitting the Rasch model to data in which the latent trait and/or the item difficulty parameters are not normally distributed may find it particularly useful. RC-IRT is also a strong candidate for cases in which the latent trait is skewed, but not when it is bimodal. In addition, RC-IRT has the advantage of being able to fit the 2PL and 3PL models, which currently cannot be done using NBE. Finally, when the latent trait or the item difficulty parameters are not normally distributed, the researcher would be best not to use estimation methods assuming normality, specifically MAP, MML, or MCMC with normal priors. Whereas these approaches performed quite well when the data conformed to the normal distribution, they each had difficulty in the various nonnormal cases, with more problems arising under greater degrees of nonnormality, particularly skewness, as has been shown in prior work.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, *7*, 473-511.

Hanson, T., & Johnson, W. O. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, *97*, 1020-1033.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Instructional Topics in Educational Measurement*, *26*(4), 38-51.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146-162. doi:10.1177/01466210122031975

Kirsci, L. & Hsu, T. C. (1995, April). *The robustness of BILOG to violations of the assumption of unidimensionality of test items and normality of ability distribution*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lavine, M. (1992). Some aspect of Polya tree distributions for statistical modeling. *Annals of Statistics*, *20*, 1222-1235.

R Core Development Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

San Martin, E., Jaran, A., Rolin, J.-M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, *76*, 385-409. doi: 10.1007/s11336-011-9213-9

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions with item response theory using true and estimated item parameters. *Applied Measurement in Education*, *21*, 65-88. doi:10.1080/08957340701796415

Seong, T. (1990) Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*, 299-311. doi:10.1177/014662169001400307

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1-16. doi:10.1177/014662169201600101

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, *11*, 253-270.

Woods, C. M. (2007). Ramsay-curve IRT for LIkert-type data. *Applied Psychological Measurement*, *31*, 195-212.

Woods, C. M. (2008). Ramsay curve item response theory for the three-parameter item response theory model. *Applied Psychological Measurement*, *36*, 447-465. doi: 10.1177/0146621607308014

Woods, C. M. (2015). Estimating the latent density in unidimensional IRT to permit non-normality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 60-84). New York, NY: Routledge.

Woods, C. M., & Linn, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*, 102-117. doi: 10.1177/0146621608319512

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281-301.