NUTS for Mixture IRT Models

Rehab Al Hakmani and Yanyan Sheng

Abstract The No-U-Turn Sampler (NUTS) is a relatively new Markov chain Monte Carlo (MCMC) algorithm that avoids the random walk behavior that common MCMC algorithms such as Gibbs sampling or Metropolis Hastings usually exhibit. Given the fact that NUTS can efficiently explore the entire space of the target distribution, the sampler converges to high-dimensional target distributions more quickly than other MCMC algorithms and is hence less computational expensive. The focus of this study is on applying NUTS to one of the complex IRT models, specifically the two-parameter mixture IRT (Mix2PL) model, and further to examine its performance in estimating model parameters when sample size, test length, and number of latent classes are manipulated. The results indicate that overall, NUTS performs well in recovering model parameters. However, the recovery of the class membership of individual persons is not satisfactory for the three-class conditions. Findings from this investigation provide empirical evidence on the performance of NUTS in fitting Mix2PL models and suggest that researchers and practitioners in educational and psychological measurement should benefit from using NUTS in estimating parameters of complex IRT models.

Keywords Markov chain Monte Carlo · No-U-Turn sampler · Mixture IRT models

1 Introduction

Classical test theory (CTT; Novick, 1966) has served the measurement community well for most of the last century. However, problems emerged using CTT have encouraged the development of a modern test theory, namely the item response theory (IRT; Lord, 1980), which has become a fundamental tool for measurement professionals

R. Al Hakmani (✉) · Y. Sheng
Southern Illinois University, 62901 Carbondale, IL, USA
e-mail: rehab.hekmani@siu.edu

Y. Sheng
e-mail: ysheng@siu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_3

in behavioral sciences (van der Linden & Hambleton, 1997). IRT consists of a family of models that specify the probability of a response given person latent trait and item characteristics. Different models exist for different types of response data. Conventional dichotomous IRT models (e.g., Birnbaum, 1969; Lord, 1980; Lord & Novick, 1968; Rasch, 1960), including the one-parameter logistic (1PL), the two-parameter logistic (2PL), and the three-parameter logistic (3PL) models, are used when test items require binary responses such as true-false questions or multiple-choice questions that are scored as correct or incorrect.

These conventional IRT models assume that the observed response data stem from a homogenous population of individuals. This assumption, however, limits their applications in test situations where, for example, a set of test items can be solved with different cognitive strategies. If the population consists of multiple groups of persons, with each group employing a different strategy for the same item, the parameters for this item will be different across these groups (or subpopulations), and consequently, the conventional IRT models cannot be used for the response data. On the other hand, the conventional IRT models may hold when each of the subpopulations employs a common strategy. As a result, mixture IRT (MixIRT; Rost, 1990) models have been developed to capture the presence of these latent classes (i.e. latent subpopulations) that are qualitatively different but within which a conventional IRT model holds. MixIRT models have become increasingly popular as a technique for investigating various issues in educational and psychological measurement such as identifying items that function differently across latent groups (e.g., Choi, Alexeev & Cohen, 2015; Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton 2002; Maij-de Meij, Kelderman, & van der Flier, 2008; Samuelsen, 2005; Shea, 2013; Wu et al., 2017) or detecting test speededness (e.g., Bolt, Cohen, & Wollack, 2002; Meyer, 2010; Mroch, Bolt, & Wollack, 2005; Wollack, Cohen, & Wells, 2003).

Over the past decades, the estimation of IRT and particularly MixIRT models has moved from the traditional maximum likelihood (ML) approach to the fully Bayesian approach via the use of Markov Chain Monte Carlo (MCMC) techniques, whose advantages over ML have been well documented in the IRT literature (e.g., de la Torre, Stark, & Chernyshenko, 2006; Finch & French, 2012; Kim, 2007; Wollack, Bolt, Cohen, & Lee, 2002). The common MCMC algorithms, such as Gibbs sampling (Geman & Geman, 1984) and Metropolis Hastings (MH; Hastings, 1970; Metropolis & Ulam, 1949), have been applied to estimate MixIRT models (e.g., Cho, Cohen, & Kim, 2013; Huang, 2016; Samuelsen, 2005; Shea, 2013). These algorithms, however, suffer from problems of inefficiently exploring the parameter space due to their random walk behavior (Neal, 1992). Recent developments of MCMC focus on non-random walk MCMCs such as the no-U-turn sampler (NUTS; Hoffman & Gelman, 2011), which can converge to high dimensional posterior distributions more quickly than common random walk MCMC algorithms, and is hence less computational expensive. In the IRT literature, Zhu, Robinson, and Torenvlied (2015) applied NUTS to simple IRT models and demonstrated its advantage over Gibbs sampling in the efficiency of the algorithm. Although NUTS has been applied with simple unidimensional IRT models (e.g., Chang, 2017; Luo & Jiao, 2017; Grant, Furr, Carpenter,

& Gelman, 2016), to date, no research has investigated its application to the more complex IRT models, such as MixIRT models.

1.1 Two-Parameter Mixture IRT Model

In the MixIRT modeling framework, persons are characterized by their location on a continuous latent dimension as well as by their latent class membership. Also, each subpopulation has a unique set of item parameters (e.g., difficulty, or discrimination). This study focuses on the two-parameter mixture (Mix2PL) IRT model, which can be viewed as an extension of the mixture Rasch model proposed by Rost (1990). If we let Y_{ij} detonate a correct ($Y_{ij} = 1$) or incorrect ($Y_{ij} = 0$) response for person i to item j, the probability of a correct response in the Mix2PL model is defined as

$$\begin{aligned} P\left(Y_{ij} = 1|\theta\right) &= \sum_{g=1}^{G} \pi_g \times P\left(Y_{ij} = 1\middle|\theta_{ig}, b_{jg}, a_{jg}, g\right) \\ &= \sum_{g=1}^{G} \pi_g \times \frac{\exp\left[a_{jg}\left(\theta_{ig} - b_{jg}\right)\right]}{1 + \exp\left[a_{jg}\left(\theta_{ig} - b_{jg}\right)\right]}, \end{aligned} \tag{1}$$

where $g = 1, \ldots, G$ is the latent class indicator, θ_{ig} denotes the ability for person i in class g, π_g denotes the proportion of examinees (i.e., the mixing proportion) in each class with a constraint that all these proportions sum to one, and b_{jg} and a_{jg} are the difficulty and discrimination parameters, respectively, for item j in the gth class.

1.2 Non-random Walk MCMC

Random walk algorithms such as Gibbs sampling and MH explore the parameter space via inefficient random walks (Neal, 1992). For complicated models with many parameters, these methods may require an unacceptably long time to converge to the target posterior distribution. On the other hand, non-random walk algorithms such as Hamiltonian Monte Carlo (HMC; Duane, Kennedy, Pendleton, & Roweth, 1987) and NUTS avoid the inefficient exploration of the parameter space. Specifically, HMC borrowed its idea from physics to suppress the random walk behavior by means of an auxiliary variable, momentum, that transforms the problem of sampling from a target posterior distribution into the problem of simulating Hamiltonian dynamics, allowing it to move much more rapidly through the posterior distribution (Neal, 2011). The unknown parameter vector $\boldsymbol{\theta}$ is interpreted as the position of a fictional particle. The Hamiltonian is an energy function for the joint state of the position $\boldsymbol{\theta}$ and the momentum $\boldsymbol{\phi}$, which defines a joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y})$. At each iteration, a random momentum vector $\boldsymbol{\phi}$ is generated, which is usually drawn from a multi-

variate normal distribution $N(\mu, \boldsymbol{\Sigma})$ with mean μ and covariance matrix $\boldsymbol{\Sigma}$. Then, the path of the particle is simulated with a potential energy equal to the negative value of the log of the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.Values of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ are simultaneously updated over time using the leapfrog algorithm, which breaks the time into discrete steps such that the total Hamiltonian simulation time is the product of the discretization interval (or the step size ε) and the number of steps taken per iteration (or the leapfrog steps L). After a Metropolis decision step is applied, the whole process repeats for an adequate number of iterations until convergence is reached (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2014; Stan Development Team, 2017).

Although HMC is a powerful MCMC technique, it requires choosing suitable values for three parameters (i.e., the step size ε, the number of leapfrog steps L, and the mass matrix $\boldsymbol{\Sigma}$) for the fictional particle. Tuning these parameters, and specifically L, requires expertise and a few preliminary runs, which can be challenging (Neal, 2011; Hoffman & Gelman, 2011). To overcome this, Hoffman and Gelman (2011) introduced NUTS to eliminate the need to set the number of leapfrog steps that the algorithm takes to generate a proposal state. Using a recursive algorithm, NUTS creates a set of candidate points that spans a wide path of the target posterior distribution, stopping automatically when it starts to double back and retrace its steps (i.e. starts to make a U-turn). Empirically, NUTS performs as efficiently as, and sometimes better than, a well-tuned HMC without requiring user interventions. Thus, NUTS is a tune-free technique, which will make it easily accessible by practitioners and researchers in behavioral sciences to fit various complex measurement models.

In view of the above, the purpose of this study is to investigate how NUTS performs in recovering parameters of the Mix2PL model under various test conditions where sample size, test length, and number of latent classes are taken into consideration. The significance of the study lies in that it not only demonstrates the application of a more efficient MCMC algorithm to the more complex MixIRT model, but also provides guidelines to researchers and practitioners on the use of such models under the fully Bayesian framework. The successful implementation of NUTS to the Mix2PL model will also help researchers with fitting more complex IRT models using fully Bayesian estimation. Findings from this investigation will provide empirical evidence and shed light on the performance of NUTS in fitting more complicated IRT models.

1.3 Model Identification

Given the difference between Bayesian and likelihood identifiability (Gelfand & Sahu, 1999), the Mix2PL IRT model was identified under the fully Bayesian approach following the literature to avoid two problems: (a) the indeterminacy and (b) the problem of label switching, which is inherent in mixture models in general. The usual practice to avoid the indeterminacy in MixIRT models, as recommended by Rost (1990), is to impose a sum-to-zero constraint in the item difficulty parameter within each latent class (i.e., $\sum_j b_{jg} = 0$). Under the fully Bayesian estimation using NUTS, there are several methods available to enforce a sum-to-zero constraint on a

parameter vector (see Stan Development Team, 2017 for more details). Due to its ease in implementation, soft centering was used in this study to apply the sum-to-zero constraint on the difficulty parameter in each latent class (i.e., $b_g \sim N(0, 1)$). Further, one practice for avoiding the problem of label switching of mixture components in MixIRT models, under the fully Bayesian framework, is to impose an ordinal constraint on the mixing proportions or the difficulty parameter (e.g., Bolt et al., 2002) or other parameters (e.g., Meyer, 2010) across latent classes. In this study, an ordinal constraint had to be imposed on both the mean ability (μ_g) parameters and the item difficulty parameters (b_g) to ensure Bayesian identifiability with Mix2PL models.

2 Methods

Monte Carlo simulations were carried out to investigate the performance of NUTS in terms of parameter recovery of the Mix2PL model under various test conditions. Data were generated using the Mix2PL model as defined in Eq. (1) with equal proportions (i.e., equal class sizes) while manipulating three factors: test length ($J = 20$ or 30), number of latent classes ($G = 2$ or 3), sample size in each subpopulation ($n =$ 250 or 500). Specifically, for the two-class condition ($G = 2$), the total number of subjects (N) was 500 or 1000; the mixing proportions were $\pi_1 = 0.50$ and $\pi_2 = 0.50$; the person ability parameters were generated from a mixture of two subpopulations where $\theta_1 \sim N(-2, 1)$ and $\theta_2 \sim N(2, 1)$; the class-specific item difficulty parameters were generated from a uniform distribution where $b_1 \sim U(-2, 0)$ and $b_2 \sim U(0, 2)$; and the class-specific item discrimination parameters were generated from a uniform distribution where $a_g \sim U(0, 2)$, $g = 1$ or 2. For the three-class condition ($G = 3$), the total number of subjects was 750 or 1500; the mixing proportions were $\pi_1 = 0.33$, $\pi_2 = 0.33$, and $\pi_3 = 0.33$; the person ability parameters were generated from a mixture of three subpopulations where $\theta_1 \sim N(-4, 1)$, $\theta_2 \sim N(0, 1)$, and $\theta_3 \sim N(4, 1)$; the class-specific item difficulty parameters were generated from a uniform distribution where $b_1 \sim U(-2, -0.5)$, $b_2 \sim U(-0.5, 0.5)$, and $b_3 \sim U(0.5, 2)$; and the class-specific item discrimination parameters were generated from a uniform distribution where $a_g \sim U(0, 2)$, $g = 1$, 2, or 3.

Priors and hyperpriors were selected to be comparable to those adopted by others (e.g., Bolt, Cohen, & Wollak, 2002; Meyer, 2010; Li, Cohen, Kim, & Cho, 2009; Wollack et al., 2003). Specifically, normal prior densities were used for person ability parameters $\theta_{ig} \sim N(\mu_g, 1)$, with a standard normal distribution for the hyperparameters μ_g, and a Dirichlet distribution for the mixing-proportion parameters such that $(\pi_1, \ldots, \pi_G) \sim \text{Dirichlet}(1, \ldots, 1)$.

Convergence of the Markov chains was examined using the Gelman-Rubin R statistic (Gelman & Rubin, 1992), with a threshold of 1.10 as suggested by Gelman et al. (2014). For the conditions involving two latent classes, the warm-up stage of either 2000 or 3000 iterations followed by 3 chains with either 3000 or 5000 sampling iterations was sufficient for the chains to reach convergence when the sample size

was 500 or 1000, respectively. For the conditions involving three latent classes, in order to reach convergence, the warm-up stage had to reach 3000, 5000 or 8000 iterations followed by 3 chains with 5000, 7000 or 10,000 sampling iterations for $N = 750$ or $N = 1000$, respectively. Ten replications were conducted for each of the simulated condition. The precision of the class and item parameter estimates was evaluated using bias and root mean square error (*RMSE*), which are defined as

$$bias_{\xi} = \frac{\sum_{r=1}^{R}\left(\widehat{\xi_r} - \xi\right)}{R}, \tag{2}$$

$$RMSE_{\xi} = \sqrt{\frac{\sum_{r=1}^{R}\left(\widehat{\xi_r} - \xi\right)^2}{R}}, \tag{3}$$

where ξ is the true value of the parameter (e.g., μ_g, π_g, a_{jg}, or b_{jg}), and $\hat{}\xi$ is the estimated value of the parameter in the rth replication where $r = 1, \ldots, R$. To summarize the recovery of item parameters, these measures were averaged over items. Further, the recovery of class memberships was evaluated by computing the percentage of correct classifications of individual persons into the class from which they were simulated. This was achieved by first calculating the probability of membership in each class g for each individual. Then, each individual was assigned to the latent class for which he or she has the highest probability of belonging (i.e., the largest membership probability).

3 Results

3.1 Mixing-Proportion and Mean Ability Recovery

The results for recovering the mixing proportion and mean ability for each latent class in the Mix2PL model are summarized in Tables 1 and 2 for the two- and three-class conditions, respectively. The small values of *bias* and *RMSE* suggest that NUTS performed well in recovering the mixing-proportion and mean ability parameters under all simulated conditions, no matter whether there were two or three latent classes. For the two-class scenarios, the *RMSE*s for estimating the mixing-proportion parameters tended to decrease with the increase of either sample size or test length. However, this pattern was not observed with the three-class scenarios or with the recovery of the mean abilities. Given that both two- and three-class conditions considered the same sample size per class ($n = 250$ or 500) and test length ($J = 20$ or 30) conditions, parameter recovery results can also be compared across the $G = 2$ versus $G = 3$ scenarios. Hence, a comparison of Tables 1 and 2 reveals that the *RMSE*s for estimating the mixing-proportion parameters tended to decrease with the increase in the number of latent classes from two to three classes, except for one scenario (i.e., $N = 1000$, $J = 30$). This is, however, not the case with

Table 1 Bias and RMSE for recovering mixing-proportion and mean ability parameters when $G = 2$

N	J	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
500	20	π_1	−0.004	0.019	μ_1	−0.016	0.205
		π_2	0.004	0.019	μ_2	−0.085	0.196
	30	π_1	−0.003	0.013	μ_1	−0.003	0.013
		π_2	0.003	0.013	μ_2	0.003	0.013
1000	20	π_1	0.005	0.012	μ_1	0.116	0.197
		π_2	−0.005	0.012	μ_2	−0.039	0.146
	30	π_1	−0.001	0.011	μ_1	0.111	0.152
		π_2	0.001	0.011	μ_2	−0.089	0.150

Table 2 Bias and RMSE for recovering mixing-proportion and mean ability parameters when $G = 3$

N	J	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
750	20	π_1	−0.002	0.012	μ_1	0.074	0.242
		π_2	0.001	0.012	μ_2	−0.008	0.102
		π_3	0.001	0.010	μ_3	−0.026	0.260
	30	π_1	−0.002	0.008	μ_1	−0.292	0.333
		π_2	−0.002	0.010	μ_2	−0.034	0.133
		π_3	0.006	0.010	μ_3	0.190	0.293
1500	20	π_1	−0.001	0.010	μ_1	0.032	0.260
		π_2	−0.001	0.008	μ_2	−0.031	0.085
		π_3	0.002	0.007	μ_3	0.075	0.189
	30	π_1	−0.005	0.007	μ_1	−0.282	0.355
		π_2	0.010	0.012	μ_2	−0.025	0.062
		π_3	−0.005	0.008	μ_3	0.284	0.363

the mean ability parameters, whose *RMSE*s tended to increase when $G = 2$ increased to $G = 3$.

It is further noted that for the three-class scenarios, the accuracy of estimating the mean ability of the second latent class was better than that of the first or third latent class (see Table 2). In addition, the precision of the mean ability estimates for the second latent class improved with the increase in the sample size.

3.2 Item Parameter Recovery

The results for recovering the difficulty and discrimination parameters are summarized in Tables 3 and 4 for the two- and three-class conditions, respectively. These

Table 3 Average *Bias* and *RMSE* for recovering item parameters when $G = 2$

N	*J*	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
500	20	a_1	−0.074	0.397	b_1	0.396	0.626
		a_2	−0.063	0.400	b_2	−0.457	0.669
	30	a_1	−0.061	0.356	b_1	0.419	0.601
		a_2	−0.055	0.359	b_2	−0.493	0.691
1000	20	a_1	−0.014	0.298	b_1	0.447	0.638
		a_2	−0.076	0.339	b_2	−0.397	0.594
	30	a_1	−0.020	0.288	b_1	0.436	0.616
		a_2	−0.037	0.300	b_2	−0.382	0.609

Table 4 Average *Bias* and *RMSE* for recovering item parameters when $G = 3$

N	*J*	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
750	20	a_1	−0.054	0.409	b_1	0.386	0.522
		a_2	−0.049	0.469	b_2	0.057	0.421
		a_3	−0.053	0.443	b_3	−0.398	0.590
	30	a_1	−0.108	0.413	b_1	0.341	0.509
		a_2	−0.078	0.482	b_2	0.017	0.396
		a_3	−0.085	0.452	b_3	−0.375	0.545
1500	20	a_1	0.023	0.339	b_1	0.352	0.517
		a_2	−0.058	0.482	b_2	0.054	0.421
		a_3	−0.096	0.419	b_3	−0.311	0.499
	30	a_1	−0.058	0.383	b_1	0.377	0.558
		a_2	−0.071	0.420	b_2	0.035	0.379
		a_3	−0.088	0.356	b_3	−0.421	0.579

results indicate that with smaller average *bias* or *RMSE*, NUTS was more accurate in recovering the discrimination parameter than the difficulty parameter of the Mix2PL model for both classes in the two-class condition and for the first and third classes in the three-class condition.

The small negative values of the average *bias* for estimating the discrimination parameters suggest that they were slightly underestimated across all the simulated conditions except for one condition (i.e., $N = 1500$ and $J = 20$) where the discrimination for the first class was overestimated (see Table 4). For the two-class condition, the recovery of the discrimination parameters improved with the increase in sample size or test length, however, this pattern was not observed in the three-class condition, which has mixed results.

The difficulty parameters were consistently underestimated for the last latent class while overestimated for the other classes, no matter whether there were two or three classes. Also for the three-class condition, the recovery of the difficulty parameters

Table 5 Percent of correct classifications of individual persons

$G = 2$					$G = 3$				
N	J	Average	Min	Max	N	Average	J	Min	Max
500	20	90.96	74.40	97.20	750	69.65	20	65.20	81.60
	30	92.38	80.80	97.20		69.91	30	66.53	87.60
1000	20	93.55	82.80	96.10	1500	71.59	20	66.60	83.40
	30	94.44	86.50	97.20		75.13	30	64.20	90.73

in the second class, as indicated by the average values of *bias* and *RMSE*, was better than the recovery of those in the first or third class across the four data sizes.

In addition, a comparison of Tables 3 and 4 suggests that the average *RMSE*s for estimating the discrimination parameter tended to increase with the increase in the number of latent classes. On the hand, the *RMSE*s for estimating the difficulty parameters tended to decrease with the increase in the number of latent classes.

3.3 Class Membership Recovery

For the class membership, the percentages of correct classifications of individual persons were computed and displayed in Table 5, which suggests that NUTS was fairly accurate when the population consisted of two latent subpopulations. The average percentages of correct classifications, across the ten replications, for the four data sizes were 90.96, 92.38, 93.55, and 94.44. However, in the conditions where the population consisted of three latent subpopulations, the recovery was less accurate, where the average percentages of correct classifications for the four data sizes were 69.65, 69.91, 71.59, and 75.13. Moreover, the recovery of class memberships is apparently affected by sample size and test length. Specifically, the average percentage of correct classifications increased with an increase in sample size or test length, for both the two- and the three-class conditions.

4 Discussion and Conclusion

With Monte Carlo simulations, results of this study suggest that overall, NUTS performs well in recovering parameters for the Mix2PL model, including the class parameters (π_g and μ_g), item parameters (a_{jg} and b_{jg}), and class membership (g), although the recovery of the class membership of individual persons is not satisfactory for the three-class condition.

With respect to the effects of sample size or test length, they play a role in recovering the class membership no matter whether the generated data sets consisted of two or three latent subpopulations. This is consistent with previous research (e.g.,

Cho et al., 2013) where the proportion of correct classification of class membership increased with either sample size or test length. However, their effects on estimating other parameters in the Mix2PL model is not clear, as some patterns of recovery improvement with the increment of sample size and/or test length in the two-class condition are not observed in the three-class condition. For example, for the two-class condition, the accuracy of estimating the mixing-proportion parameters increases with the increase of either sample size or test length but this pattern is not observed with the three-class condition. This is possibly due to the increased complexity of the mixture item response theory (MixIRT) model with the increased number of latent classes. Adding one subpopulation may seem trivial, but it would result in a substantial increase in the number of parameters to be estimated. This complexity is further reflected in the estimation of person mean ability or item discrimination parameters, whose accuracy decreases with the increased number of classes. On the other hand, the recovery of the mixing proportions or individual item difficulties in the model is not seemingly affected by such added complexity. As a matter of fact, their *RMSE* values decrease when adding one more subpopulation. This reduction is due to the fact that the magnitude of *RMSE* depends on the unit/scale of the parameter. For instance, the mixing proportion is larger for the two-class condition ($\pi_g = 0.5$) than the three-class condition ($\pi_g = 0.33$), and hence the *RMSEs* tend to be larger with the two-class condition. This is certainly a limitation of using *RMSE* for evaluating the accuracy in recovering model parameters in this study. Future studies shall consider other measures, such as the relative *RMSE* or normalized *RMSE* that are free from the scale of the parameters.

The finding that the discrimination parameter is better recovered than the difficulty parameter in the MixIRT model (based on the comparison of average *RMSE*/*bias* values) agrees with Chang (2017), who focused on the estimation of the conventional IRT model using NUTS and Gibbs sampling. However, it does not agree with findings from studies on fitting some other IRT models with non-Bayesian estimations (e.g., Batley & Boss, 1993; Kang & Cohen, 2007) although the same *RMSE* criterion has been used. Given the limitation of *RMSE* as noted previously, further studies are needed to direct the trend of such comparisons. In addition, results based on the three-class situation suggest that the item difficulty or the class mean ability parameters are estimated more accurately for the second class than for the first or third class. This is likely due to the choice of the simulated person ability and item difficulty parameters for each of the three latent classes. Specifically, the generated person abilities for the second class (i.e., $\theta_2 \sim N(0, 1)$) coincides with the generated item difficulty (i.e., $b_2 \sim U(-0.5, 0.5)$) for that class. However, the generated person abilities for the first class (i.e., $\theta_1 \sim N(-4, 1)$) is quite distant from the generated item difficulty (i.e., $b_2 \sim U(-2, -0.5)$) for that class, such that the average person ability (i.e., -4) is 2.75 standard deviations lower than the average item difficulty (i.e., -1.25). Similarly, the generated person ability for the third class (i.e., $\theta_3 \sim N(4, 1)$) is quit distant from the generated item difficulty (i.e., $b_2 \sim U(0.5, 2)$) for that class, such that the average person ability (i.e., 4) is 2.75 standard deviations higher than the average item difficulty (i.e., 1.5). Thus, in order to obtain more accurate estimates of the person mean ability and item difficulty parameters for the first class,

more easy items should be added. On the other hand, in order to obtain more precise estimations of the person mean ability and item difficulty parameters for the third class, more difficult items should be added.

This study provides empirical evidence on the performance of NUTS in fitting MixIRT models. It also shows that researchers and practitioners in educational and psychological measurement can use NUTS in estimating parameters of complex IRT models such as MixIRT models. However, conclusions that are made in the present study are based on the simulated conditions and cannot be generalized to other conditions. Therefore, for future studies, additional test conditions need to be explored such as unequal mixing proportions, small sample size, and short test length. Given the computational expense of fitting NUTS to the complex Mix2PL model, this study only used 10 replications for each experimental condition. However, as suggested by Harwell, Stone, Hsu, & Kirisci, (1996), a minimum of 25 replications is recommended for typical Monte Carlo studies in IRT modeling. Additional studies with similar experimental conditions are needed before one can conclude about the use of the algorithm with fitting the Mix2PL model and further the effects of sample size, test length, and number of classes on estimating the model. In addition, this study focused on the dichotomous Mix2PL model. Future studies may consider evaluating the performance of NUTS using other dichotomous MixIRT models such as the Mix1PL model or the Mix3PL models, or using MixIRT models with polytomous categories such as a mixture version of Bock's (1972) nominal response model or a mixture version of Masters's (1982) partial credit model. Moreover, this study considered certain population distributions and difficulty ranges. Additional studies are necessary to consider other person distributions and/or other ranges for item difficulty parameters to decide on the test condition that leads to more accurate estimates for all classes. Future studies are also needed to decide on the optimal number of persons and/or items for more accurate estimations of class membership in conditions where the population includes three or more subpopulations for any given class size. Finally, findings from this study are based on simulated conditions where the true parameters are known, Future studies may adopt NUTS algorithms to fit the Mix2PL models to real data and examine how NUTS performs in real test situations.

References

Batley, R.-M., & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement, 17*(2), 131–141. https://doi.org/10.1177/014662169301700203.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*(2), 258–276.

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331–348.

Chang, M. (2017). *A comparison of two MCMC algorithms for estimating the 2PL IRT models*. Doctoral: Southern Illinois University.

Cho, S., Cohen, A., & Kim, S. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83*(2), 278–306.

Choi, Y., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing, 15*(3), 239–253. https://doi.org/10.1080/15305058.2015.1007241.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement Summer, 42*(2), 133–148.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: a mixture distribution conceptualization. *International Journal of Testing, 2*(3&4), 243–276.

de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement, 30*(3), 216–232. https://doi.org/10.1177/0146621605282772.

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B, 195,* 216–222. https://doi.org/10.1016/0370-2693(87)91197-X.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods, 11*(1), 167–178.

Gelfand, A. E., & Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *JASA, 94*(445), 247–253. https://doi.org/10.2307/2669699.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Florida: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat Sci, 7*(4), 457–472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

Grant, R. L., Furr, D. C., Carpenter, B., & Gelman, A. (2016). Fitting Bayesian item response models in Stata and Stan. *The Stata Journal, 17*(2), 343–357. https://arxiv.org/abs/1601.03443. Accessed 18 Apr 2018.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101–125. https://doi.org/10.1177/014662169602000201.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97.

Hoffman, M. D., & Gelman, A. (2011). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(2), 1593–1624.

Huang, H. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01706.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331–358. https://doi.org/10.1177/0146621606292213.

Kim, S.-H. (2007). Some posterior standard deviations in item response theory. *Educational and Psychological Measurement, 67*(2), 258–279. https://doi.org/10.1177/0013164407067002050l.

Li, F., Cohen, A., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*(5), 353–373. https://doi.org/10.1177/0146621608326422.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (2nd ed.). New Jersey: Hillsdale.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Maryland: Addison-Wesley.

Luo, Y., & Jiao, H. (2017). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, 1–25. https://doi.org/10.1177/0013164417693666.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research, 45*(6), 975–999. https://doi.org/10.1080/00273171.2010.533047.

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

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 44*(247), 335–341.

Meyer, J. P. (2010). A mixture Rasch model with Item response time components. *Applied Psychological Measurement, 34*(7), 521–538. https://doi.org/10.1177/0146621609355451.

Mroch, A. A., Bolt, D. M., & Wollack, J. A. (2005). *A new multi-class mixture Rasch model for test speededness*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebe, April 2005.

Neal, R. M. (1992). *An improved acceptance procedure for the hybrid Monte Carlo algorithm*. Retrieved from arXiv preprint https://arxiv.org/abs/hep-lat/9208011.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162). Florida: CRC Press.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (2nd ed.). Danmark: Danmarks Paedagogiske Institute.

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

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Dissertation). University of Maryland.

Shea, C. A. (2013). *Using a mixture IRT model to understand English learner performance on large-scale assessments* (Dissertation). University of Massachusetts.

Stan Development Team. (2017). *Stan modeling language users guide and reference manual, version 2.17.0*. http://mc-stan.org. Accessed 8 Feb 2018.

van der Linden, Wd, & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: a comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*(3), 339–352. https://doi.org/10.1177/0146621602026003007.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40,* 307–330.

Wu, X., Sawatzky, R., Hopman, W., Mayo, N., Sajobi, T. T., Liu, J., … Lix, L. M. (2017). *Latent variable mixture models to test for differential item functioning: a population-based analysis*. *Health and Quality of Life Outcomes, 15*. https://doi.org/10.1186/s12955-017-0674-0.

Zhu, L., Robinson, S. E., & Torenvlied, R. (2015). A Bayesian approach to measurement bias in networking studies. *The American Review of Public Administration, 45*(5), 542–564. https://doi.org/10.1177/0275074014524299.