International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 568–579
https://dx.doi.org/10.21449/ijate.581314
Published at http://www.ijate.net http://dergipark.org.tr *Research Article*

The Uniform Prior for Bayesian Estimation of Ability in Item Response Theory Models

Tuğba Karadavut [1,*]

[1] Recep Tayyip Erdogan University, Faculty of Education, Cayeli, Rize, Turkey

ARTICLE HISTORY

Received: 22 June 2019
Revised: 16 September 2019
Accepted: 17 October 2019

KEYWORDS

Item response theory,
Uniform ability,
Bayesian estimation

Abstract: Item Response Theory (IRT) models traditionally assume a normal distribution for ability. Although normality is often a reasonable assumption for ability, it is rarely met for observed scores in educational and psychological measurement. Assumptions regarding ability distribution were previously shown to have an effect on IRT parameter estimation. In this study, the normal and uniform distribution prior assumptions for ability were compared for IRT parameter estimation when the actual distribution was either normal or uniform. A simulation study that included a short test with a small sample size and a long test with a large sample size was conducted for this purpose. The results suggested using a uniform distribution prior for ability to achieve more accurate estimates of the ability parameter in the 2PL and 3PL models when the true distribution of ability is not known. For the Rasch model, an explicit pattern that could be used to obtain more accurate item parameter estimates was not found.

1. INTRODUCTION

Item Response Theory (IRT) is widely used in psychological measurement (Embretson, 1996), and in educational measurement (Lord & Novick, 1968) for designing and analyzing the measurement instruments. It also has applications in other fields such as public health, ecology and sociology. In educational measurement, the student ability is the subject of the measurement. Ability is a latent trait and it cannot be measured directly. Thus, student responses to items in a test are used to measure ability in educational measurement. IRT defines a continuous and monotonic mathematical function (Reckase, 2009) for explaining the relationship between latent ability and student responses to the test items (Embretson & Reise, 2000). In this study, the latent ability is assumed to be unidimensional, and the IRT models that are for analyzing the unidimensional latent ability are considered for the analysis.

The estimation methods for IRT models require an assumption regarding the ability distribution to enable estimation of the model parameters. The tradition is to assume a normal distribution for abilityfor estimating the model parameters. Generally, normality is a reasonable assumption for ability (Embretson & Reise, 2000). However, it is not unlikely for observed scores in educational and psychological measurement to be non-normal in reality (e.g., Cook, 1959; Lord, 1955; Micerri, 1989). Micerri (1989), in example, examined 440 raw-score distributions

CONTACT: Tuğba Karadavut ✉ tugba-mat@hotmail.com Recep Tayyip Erdogan University, Faculty of Education, Cayeli, Rize, Turkey

ISSN-e: 2148-7456 /

from large-scale achievement and psychometric measures. Micerri (1989) found that, of the measures he investigated, 125 were moderately asymmetric (i.e., 28.4%), and 135 were extremely asymmetric (i.e., 30.7%). The non-normality in the observed scores may also indicate non-normality in the latent ability scores. It is because the observed raw scores and the latent ability scores from an IRT model are correlated (Fan, 1998; Stewart, 2012).

There are two general methods for estimation of the parameters in IRT models. These are marginal maximum likelihood estimation (Bock & Aitkin, 1981) and Bayesian estimation methods. Both of these methods make prior assumptions regarding the ability distribution (Baker & Kim, 2004, de Ayala, 2009). In this study, Markov chain Monte Carlo (MCMC) estimation was used for estimation of the model parameters. MCMC is a Bayesian estimation technique that iteratively samples from the posterior distributions of the parameters to be estimated (Jackman, 2000). These samples are then used to obtain estimates of the parameters. Bayesian estimation methods require indication of a prior distribution for each parameter in the model that is intended to be estimated. The prior distribution for a parameter reflects the distributional assumptions regarding that parameter. Poor specification of the priors in Bayesian estimation may result in biased parameter estimates (e.g., Mislevy, 1986). Therefore, a sufficiently informative prior should be specified for each parameter in the model in order to obtain unbiased estimates of the parameters (Baker & Kim, 2004; Mislevy, 1986). A sufficiently informative prior provides information regarding the posterior distribution of the parameter to be estimated. The prior may be assumed to be from the same distribution family with the posterior distribution (e.g., conjugate prior).

Assumptions with respect to ability distribution have been shown to have an effect on IRT parameter estimation, depending on the deviation from the actual ability distribution (Reise & Yu, 1990; Roberts, Donoghue, & Laughlin, 2002; Sass, Schmitt, & Walker, 2008; Sen, Cohen, & Kim, 2016; Seong, 1990; Stone, 1992). Item parameter estimates are more precise when the prior distribution for latent ability matches the true distribution of latent ability (Seong, 1990). The bias in item parameter estimates due to misspecification of actual ability distribution, on the other hand, often can be reduced by increasing sample size and test length (e.g., de Ayala & Sava-Bolesta, 1999; Kirisci, Hsu, & Yu, 2001, Reise & Yu, 1990; Roberts et al., 2002; Seong, 1990; Stone, 1992). Thus, the effect of the prior distributional assumptions on parameter estimation should be considered with respect to the potentially confounding variables such as the sample size and the test length.

The latent ability distribution in an IRT model can be estimated with an assumption of normality following the general applications in the literature. The true ability distribution, on the other hand, can be in another type such as the uniform distribution (e.g., Hambleton & Cook, 1983; Swaminathan, Hambleton, & Rogers, 2007). In that case, using a prior distribution that matches the true distribution of the latent ability may result in more accurate estimates of item and ability parameters in IRT models. In this study, the normal prior distribution for ability was investigated for its efficiency to result in reasonable estimates of item and ability parameters, especially when the true latent ability distribution was uniform. A simulation study was conducted to analyze student responses to items with a normal and a uniform underlying ability distribution. The analyses were done using a unidimensional IRT model for dichotomous items. The models used in this study were Rasch (Rasch, 1960), two-parameter logistic (2PL; Birnbaum, 1968), and three-parameter logistic (3PL; Birnbaum, 1968) IRT models. Uniform and normal distributions priors were used for the latent ability while analyzing student responses to items. Finally, item and ability parameter estimates from the models with a normal and a uniform prior distribution for the latent ability were compared to the generating item and ability parameters in order to determine the accuracy of item and ability parameter estimates.

2. METHOD

2.1. Unidimensional Item Response Theory Models

Unidimensional IRT models for dichotomous items (e.g., for multiple choice) are extensively used in educational measurement. These models include Rasch, 2PL and 3PL models. The names of 2PL and 3PL models vary depending on the number of item parameters in the model. Namely, the 2PL model has two item parameters that are item difficulty and item discrimination parameters. Similarly, the 3PL model includes three item parameters that are item difficulty, item discrimination and the item pseudo-guessing parameters. The 3PL model defines the probability that an examinee j with ability θ answers item i correctly ($P_i(\theta_j|X=1)$) with the following equation:

$$P_i(\theta_j|X=1) = c_i + (1-c_i)\frac{1}{1+e^{-a_i(\theta_j-b_i)}}, \quad (1)$$

where b_i is the item difficulty parameter for item i, a_i is the item discrimination parameter for item i, and c_i is the pseudo-guessing parameter for item i. Fixing the c_i parameter in a 3PL model to zero results the 3PL model to reduce into a 2PL model. Thus, the probability that an examinee j with ability θ answers item i correctly in a 2PL model is:

$$P_i(\theta_j|x=1) = \frac{1}{1+e^{-a_i(\theta_j-b_i)}}. \quad (2)$$

Similary, fixing the c_i parameter to zero and the a_i parameter to one in a 3PL model yields a Rasch model. The Rasch model defines the probability that an examinee j with ability θ answers item i correctly as:

$$P_i(\theta_j|x=1) = \frac{1}{1+e^{-(\theta_j-b_i)}}. \quad (3)$$

2.2. The Simulation Design

Binary student responses to test items were generated using the R (2016) software for the Rasch, 2PL and 3PL models. The underlying latent ability distributions were simulated to follow either a standard normal distribution or a uniform distribution on the interval [-3, 3]. Two test lengths (15-item and 30-item) and two sample sizes (600 and 2,000) were generated. Twenty-five data sets were simulated for each simulation condition. Item parameters that are used to generate student responses to test items are given in Table 1.

Table 1. Item Parameter Estimates Used for Generating Student Responses.

	Rasch	2PL		3PL		
	b	***b***	***a***	***b***	***a***	***c***
1	2.75	2.75	1.0	2.75	1.0	0.25
2	2.50	2.50	1.0	2.50	1.0	0.25
3	2.25	2.25	1.0	2.25	1.0	0.25
4	2.00	2.00	1.0	2.00	1.0	0.25
5	1.75	1.75	1.0	1.75	1.0	0.25
6	1.50	1.50	1.5	1.50	1.5	0.15
7	1.25	1.25	1.5	1.25	1.5	0.15
8	1.00	1.00	1.5	1.00	1.5	0.15
9	0.75	0.75	1.5	0.75	1.5	0.15
10	0.50	0.50	1.5	0.50	1.5	0.15
11	0.25	0.25	2.0	0.25	2.0	0.10
12	0.00	0.00	2.0	0.00	2.0	0.10
13	-0.25	-0.25	2.0	-0.25	2.0	0.10
14	-0.50	-0.50	2.0	-0.50	2.0	0.10
15	-0.75	-0.75	2.0	-0.75	2.0	0.10

2.3. Estimation of the Parameters

Estimation of the parameters was done by using the Markov Chain Monte Carlo (MCMC) method as implemented in the computer software OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). A burn-in period of 3,000 iterations was used with a total number of 30,000 iterations for each model. Following priors were used for MCMC estimation of model parameters:

$$\begin{aligned} b_i &\sim \text{Normal}(0,1), \qquad i = 1, \dots, n, \\ a_i &\sim \text{Normal}(0,1) \text{ and } a_i > 0\,, \; i = 1, \dots, n, \\ c_i &\sim \text{Beta}(5,17) \text{ and } 0 < c_i < 0.3, \quad i = 1, \dots, n. \end{aligned} \tag{4}$$

Following priors were used for estimation of ability parameter, depending on the prior assumptions regarding the ability:

$$\theta_j \sim \text{Normal}(0,1), \qquad j = 1, \dots, N,$$

or

$$\theta_j \sim \text{Uniform}(-4,4), \qquad j = 1, \dots, N. \tag{5}$$

The scale of ability is arbitrary in IRT estimation which is denoted as metric identification problem (de Ayala, 2009, p. 41; Baker & Kim, 2004). The metric of the ability requires to be identified to achieve comparable parameter estimates across different calibrations. In this study, the metric of the ability was identified using item centering method (de Ayala, 2009). That is, the mean of item difficulty parameter estimates were fixed to zero for estimation of each model. In addition, the scale of parameters from estimated models was placed on scale of the generating parameters by using mean and sigma equating method (Marco, 1977).

2.4. Item Recovery Analyses

Item recovery analyses were conducted to compare the generating parameters to the parameter estimates from the MCMC analyses with a normal prior and the MCMC analyses with a uniform prior. Accuracy indices and Pearson correlations were calculated for this purpose. The accuracy indices included mean bias, mean absolute error (MAE), mean-square error (MSE), and root-mean-square error (RMSE). The mean bias, MAE, MSE, RMSE and Pearson

correlation values were calculated across twenty-five replications for the 15-item and 600 sample size condition, and for the 30-item and 2,000 sample size condition, individually, for each IRT model. As an example, the equations for calculating the accuracy indices and Pearson correlation for the item difficulty parameter (b) are given below:

$$\text{Bias}(\hat{b}) = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n}(\hat{b}_i - \hat{b}_{ir})}{Rxn}, \tag{6}$$

$$\text{MAE}(\hat{b}) = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n}|\hat{b}_i - \hat{b}_{ir}|}{Rxn}, \tag{7}$$

$$\text{MSE}(\hat{b}) = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n}(\hat{b}_i - \hat{b}_{ir})^2}{Rxn}, \tag{8}$$

$$\text{RMSE}(\hat{b}) = \sqrt{\frac{\sum_{r=1}^{R}\sum_{i=1}^{n}(\hat{b}_i - \hat{b}_{ir})^2}{Rxn}}, \tag{9}$$

$$\text{Cor}(\hat{b}, b) = \frac{1}{R}\sum_{r=1}^{R}\text{Cor}(\hat{b}_i, \hat{b}_{ir}), \tag{10}$$

where ($\hat{b}_i$) is the generating item difficulty parameter for item i, ($\hat{b}_{ir}$) is the item difficulty parameter estimate for item i from MCMC analyses with a uniform/normal prior from the rth replication, R is total number of replications which is 25, and n is the total number of items which is either 15 or 30.

3. RESULT / FINDINGS

The accuracy indices and the correlation coefficients are calculated for the item difficulty, item discrimination, item pseudo-guessing, and ability to quantify the item parameter recovery (see Appendix A, Tables A1-A4). Post-hoc comparisons were conducted for transformed MSE values using Tukey's HSD procedure (see Table 2). Square-root or natural logarithm transformation was used for transformation of the MSE values in order to achieve normally distributed residuals. Cohen's d values for the post-hoc comparisons are reported in Table 2. Cohen's d values of 0.2, 0.5, and 0.8 indicate small, medium, and large effects, respectively (Cohen, 1988). Cohen's d values of 0.8 and larger were considered to reveal a substantial difference in mean MSE values between the uniform and the normal priors for a given parameter from a particular model for a given number of the items and the sample size condition.

Results did not indicate a difference in the mean MSE values between the normal and the uniform priors for the item difficulty parameter from the Rasch model, for both the 15-item and 600 sample size and for the 30-item and 2,000 sample size conditions. There was not a constant pattern for differences in the mean MSE values between the uniform and the normal priors for the ability parameter from Rasch model.

For the 15-item and 600 sample size conditiom, there was not a substantial difference in the mean MSE values between the uniform and the normal priors for estimation of the item difficulty and the item discrimination parameters using a 2PL model, when the actual distribution of the latent ability was uniform. When the actual distribution of the latent ability was normal, the uniform prior yielded larger mean MSE value compared to the normal prior for both of the item difficulty and item discrimination parameters. For the 30-item and 2,000 sample size condition, for both of the item difficulty and item discrimination parameters, the

normal prior yielded larger mean MSE value when the actual distribution of the latent ability was uniform. Similarly, the uniform prior yielded larger mean MSE value when the actual distribution of the latent ability was normal, for both of the item difficulty and item discrimination parameters. For estimation of the ability parameter using a 2PL model, the normal prior yielded larger mean MSE values compared to the uniform prior for all conditions.

The analyses of the 15-items using a 3PL model for 600 sample size showed that, there was not a substantial difference in the mean MSE values between the uniform and the normal priors for the item difficulty and the item discrimination parameters, when the actual distribution of the latent ability was uniform. For the item pseudo-guessing parameter, the uniform prior yielded larger errors compared to the normal prior, when the actual latent ability distribution was uniform, for the 15-item and 600 sample size condition. Again for the 15-item and 600 sample size condition, the uniform prior yielded larger errors compared to the normal prior, when the actual latent ability distribution was normal, for estimation of the item difficulty, item discrimination and item pseudo-guessing parameters.

For the 30-item and 2,000 sample size condition, the normal prior yielded larger mean MSE values compared to the uniform prior, for estimation of the item difficulty and item pseudo-guessing parameters, when the actual latent ability distribution was uniform. For the item discrimination parameter, on the other hand, there was not a significant difference in the mean MSE values between the normal and uniform priors. Again for the 30-item and 2,000 sample size condition, the uniform prior yielded larger mean MSE values for the item difficulty, item discrimination, and item pseudo-guessing parameters, when the actual distribution of the latent ability was normal. For estimation of the ability parameter in the 3PL model, the normal prior yielded larger mean MSE values compared to the uniform prior, when the actual latent ability distribution was uniform. The effect sizes for the difference between the normal and uniform priors were medium to large (i.e., between 0.5 and 0.8) when the actual distribution was normal.

Table 2. Estimates of Cohen's d Values from Post-hoc Comparisons Using Tukey'd HSD Procedure for Transformed MSE Values

			Rasch		**2PL**			**3PL**			
Condition	**Actual Dist.**	**Prior Dist.**	***b***	***θ***	***b***	***a***	***θ***	***b***	***a***	***c***	***θ***
15-item and 600 sample size	Uniform	Normal – Uniform	0.138 U>N	**1.240 U>N**	0.579 N>U	0.236 U>N	**3.467 N>U**	0.611 N>U	0.780 U>N	**0.915 U>N**	**7.627 N>U**
	Normal	Normal – Uniform	0.026 U>N	0.455 U>N	**1.243 U>N**	**2.819 U>N**	**7.526 N>U**	**2.299 U>N**	**5.090 U>N**	**2.319 U>N**	0.756 N>U
30-item and 2,000 sample size	Uniform	Normal – Uniform	0.245 U>N	**3.809 U>N**	**0.999 N>U**	**1.314 N>U**	**3.197 N>U**	**2.240 N>U**	0.281 U>N	**1.466 N>U**	**6.022 N>U**
	Normal	Normal – Uniform	0.013 U>N	**2.092 N>U**	**1.231 U>N**	**3.914 U>N**	**3.903 N>U**	**2.998 U>N**	**7.894 U>N**	**1.056 U>N**	0.727 U>N

Note. 1) Dist: Distribution, N: Mean parameter estimates for the model with normal prior, U: Mean parameter estimates for the model with uniform prior, *b*: Item difficulty, *a*: Item discrimination, *c*: Item pseudo-guessing, *θ*: Ability 2) Large effect sizes (i.e., larger than .80) are shown in bold.

4. DISCUSSION and CONCLUSION

The primary purpose of using IRT models is to locate students on a continuous scale by estimating their ability (Baker, 2001). Thus, correct estimation of the ability parameters in an IRT model is critical for accountability. The purpose of this study was to investigate if the prior distribution assumption for ability has an effect on estimation of the ability parameters, especially when the true ability distribution is uniform. For this purpose, a simulation study was

conducted to compare a uniform and a normal prior distribution assumption for Bayesian estimation of the item and ability parameters in Rasch, 2PL and 3PL models. The simulation conditions included a short test with a small sample size, and a long test with a large sample size. Ability distributions were generated to follow either a normal or a uniform distribution; and the item responses were generated to fit either a Rasch, 2PL or a 3PL model. Twent-five data sets of item responses were generated for each combination of the simulation conditions. Each data set was analyzed using both a uniform and a normal prior, and the ability and item parameter estimates from both models were compared for their accuracy.

Uniform and normal priors for ability yielded similar item parameter estimates for the Rasch model for each simulation conditions. The uniform and normal priors either resulted similar item parameter estimates, or the prior that does not match the true distribution resulted in better estimates of the ability parameter. That is, a uniform distribution prior yielded more accurate estimates of the ability parameter when the true distribution of ability was normal; and the normal prior resulted in more accurate ability parameter estimates when the true distribution of ability was uniform.

For the 2PL model, the normal and the uniform distribution priors for ability either resulted in similar item difficulty and item discrimination parameter estimates, or the prior distribution that matches the true distribution of ability resulted in more accurate estimates of similar item difficulty and item discrimination parameters. For estimation of the ability parameters, the uniform distribution prior yielded more accurate estimates for each of the simulation condition independent of the true distribution of ability.

The uniform and normal distribution priors for ability either resulted in similar item difficulty parameter estimates, or the prior that matches the true distribution of ability yielded more accurate item difficulty parameter estimates. Uniform and normal distribution priors resulted in similar item discrimination parameter estimates when the true distribution was uniform. Normal distribution prior yielded more accurate estimates of the item discrimination parameter when the true distribution of ability was normal. Similarly, the normal distribution prior yielded more accurate estimates of the item pseudo-guessing parameter for each of the simulation conditions except for the 30-item and 2,000 sample size condition, when the true distribution of ability was uniform. For this condition, the uniform distribution prior resulted in more accurate estimates of the item pseudo-guessing parameters. The uniform and normal distribution priors for ability yielded similar estimates of ability when the true distribution of ability was normal. When the true distribution of ability was uniform, on the other hand, the uniform distribution prior yielded more accurate estimates of the ability parameter.

In summary, the results of this study suggest using a uniform distribution prior to achieve more accurate estimates of the ability parameter in the 2PL and 3PL models when the true distribution of ability is not known. The results contribute to the IRT literature as they suggest that using a uniform prior for ability may be more useful as opposed to the convention of using a normal prior for estimation of ability. The results did not indicate a guiding pattern for estimation of the ability parameter in the Rasch model. However, the results of this study are limited with the simulation conditions used in the study. A future study may include more alternatives for the test length and the sample size conditions. In addition, this study only investigated the effect of the prior distribution for ability on estimation of the parameters in IRT models. A future study may explore potential effects of the prior distributions for the item parameters on parameter estimation in IRT models.

Acknowledgements

An earlier version of this paper was presented at the International Conference on Research in Education and Science (ICRES), Kusadasi, Aydin, Turkey, May 18-21, 2017.

ORCID

Tuğba Karadavut https://orcid.org/0000-0002-8738-7177

5. REFERENCES

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland. Retrieved from http://files.eric.ed.gov/fulltext/ED458219.pdf

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley. Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, D. L. (1959). A replication of Lord's study on skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement, 19*, 81-87.

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

de Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*, 3-19.

Embretson, S. E. (1996). The new rules of measurement. Psychological Assessment, 8, 341.

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.

Hambleton, R. K., & Cook, L. L. (1983). *The robustness ofitem response models and the effects of test length and sample size on the precision of ability estimates*. In D. Weiss (Ed.), *New horizons in testing* (pp. 31–49). NewYork: Academic Press.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science, 44*, 375-404.

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.

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement, 15*, 383-389.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores* (with contributions by A. Birnbaum). Reading, MA: Addison-Wesley.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine, 28*, 3049-3082.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Micerri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved January 10, 2017, from https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielson and Lydiche (for Danmarks Paedagogiske Institut).

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26*, 192-207.

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

Sen, S., Cohen, A. S., & Kim, S.-H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement, 40*, 98-113.

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.

Stewart, J. (2012) Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination. *Shiken Research Bulletin, 16*, 15-22.

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

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). *Assessing the fit of item response theory models*. In C. R. Rao & S. Sinharay (Eds.), Psychometrics: Vol. 26. Handbook of statistics (pp. 683–718). Amsterdam: Elsevier.

6. APPENDIX

Table A1. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	1.037	0.977	0.000	0.000	0.036	0.030	0.900	0.932
MAE	0.083	0.086	1.078	1.011	0.073	0.080	0.124	0.114	0.916	0.949
MSE	0.011	0.012	1.506	1.405	0.009	0.011	0.025	0.022	1.064	1.145
RMSE	0.107	0.109	1.227	1.185	0.092	0.107	0.159	0.149	1.031	1.070
Cor.	0.995	0.995	0.931	0.930	0.996	0.995	0.945	0.961	0.957	0.953

Table A1 Continues. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	-0.123	-0.151	0.001	-0.018	0.922	1.039
MAE	0.124	0.136	0.234	0.205	0.035	0.030	0.978	1.059
MSE	0.026	0.033	0.078	0.061	0.002	0.001	1.295	1.477
RMSE	0.162	0.183	0.280	0.246	0.043	0.038	1.138	1.215
Cor.	0.989	0.986	0.869	0.953	0.735	0.868	0.933	0.930

Table A2. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	0.848	0.901	0.000	0.000	-0.172	-0.047	0.862	0.978
MAE	0.078	0.078	0.918	0.921	0.147	0.104	0.257	0.154	0.876	0.983
MSE	0.010	0.010	1.144	1.110	0.033	0.021	0.099	0.038	0.969	1.145
RMSE	0.099	0.099	1.069	1.054	0.182	0.145	0.314	0.195	0.985	1.070
Cor.	0.996	0.996	0.833	0.835	0.986	0.991	0.790	0.900	0.900	0.908

Table A2 Continues. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	-0.114	-0.150	0.020	0.001	0.967	1.062
MAE	0.230	0.136	0.447	0.210	0.049	0.040	1.007	1.070
MSE	0.075	0.035	0.288	0.061	0.004	0.002	1.389	1.407
RMSE	0.274	0.187	0.537	0.247	0.061	0.048	1.179	1.186
Cor.	0.968	0.985	0.226	0.913	0.432	0.656	0.868	0.875

Table A3. Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	1.136	1.000	0.000	0.000	-0.002	0.037	1.003	1.029
MAE	0.048	0.050	1.145	1.005	0.046	0.057	0.063	0.079	1.006	1.034
MSE	0.004	0.004	1.548	1.249	0.004	0.006	0.007	0.011	1.156	1.235
RMSE	0.059	0.062	1.244	1.117	0.060	0.075	0.081	0.104	1.075	1.112
Cor.	0.998	0.998	0.958	0.960	0.998	0.998	0.982	0.985	0.975	0.971

Table A3 Continues: Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	-0.172	-0.036	-0.002	-0.019	0.987	1.043
MAE	0.076	0.113	0.207	0.188	0.019	0.025	0.999	1.049
MSE	0.010	0.023	0.059	0.052	0.001	0.001	1.222	1.357
RMSE	0.099	0.153	0.242	0.229	0.025	0.033	1.106	1.165
Cor.	0.996	0.990	0.928	0.973	0.919	0.910	0.958	0.953

Table A4. Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**
Bias	0	0	0.838	0.945	0	0	-0.117	-0.014	0.945	1.003
MAE	0.047	0.047	0.863	0.948	0.090	0.061	0.158	0.076	0.946	1.003
MSE	0.003	0.003	0.935	1.071	0.013	0.009	0.037	0.009	1.023	1.115
RMSE	0.059	0.059	0.967	1.035	0.115	0.093	0.191	0.095	1.011	1.056
Cor.	0.999	0.999	0.904	0.906	0.994	0.996	0.956	0.976	0.935	0.944

Table A4 Continues. Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**
Bias	0.000	0.000	0.032	-0.032	0.025	0.005	0.932	0.968
MAE	0.182	0.088	0.470	0.145	0.036	0.029	0.941	0.969
MSE	0.047	0.016	0.364	0.034	0.002	0.001	1.101	1.087
RMSE	0.216	0.126	0.604	0.185	0.043	0.038	1.049	1.043
Cor.	0.980	0.993	-0.058	0.922	0.829	0.806	0.910	0.921