ISRES Publishing

The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational & Social Sciences (EPESS), 2017

Volume 7, Pages 30-37

ICRES 2017: International Conference on Research in Education and Science

ESTIMATION OF ITEM RESPONSE THEORY MODELS WHEN ABILITY IS UNIFORMLY DISTRIBUTED

Tugba Karadavut
Kilis 7 Aralık University, Kilis, TURKEY
The University of Georgia, Athens, GA, USA

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 assumptions for ability were compared for IRT parameter estimation, when the actual distribution was either normal or uniform. Uniform distribution assumption in 2PL model yielded more accurate estimates of ability independent of the actual ability distribution. Similarly, a uniform distribution assumption for ability yielded more accurate estimates of ability in 3PL model when the actual ability distribution was uniform. For Rasch model, there was not an explicit pattern for comparing accuracy of ability estimates from uniform and normal distribution assumptions.

***Keywords*:** Item response theory, IRT, uniform distribution, normal distribution, ability.

Introduction

Item Response Theory (IRT), also known as latent trait theory, is a modern mental test paradigm which is extensively used in psychological measurement (Embretson, 1996), and in educational measurement (Lord & Novick, 1968). It also has a wide usage in other fields such as public health, ecology and sociology. Student ability as a latent trait is unobservable and can not be measured directly. IRT defines a continuous and monotonic mathematical function (Reckase, 2009) for explaining the relationship between latent ability and student responses to test items (Embretson & Reise, 2000).

IRT models traditionally assume a normal distribution for ability while estimating ability of students. Normality is often a reasonable assumption for ability (Embretson & Reise, 2000). However, the normality assumption is rarely met for observed scores in educational and psychological measurement (e.g., Cook, 1959; Lord, 1955; Micerri, 1989). Micerri (1989), in example, examined 440 raw-score distributions from large-scale achievement and psychometric measures. Of these measures, 125 were moderately asymmetric (i.e., 28.4%), and 135 were extremely asymmetric (i.e., 30.7%).

Estimation of parameters in IRT models can be done by employing either marginal maximum likelihood estimation (Bock & Aitkin, 1981) or Bayesian estimation methods. Both methods make prior assumptions regarding the ability distribution (Baker & Kim, 2004, de Ayala, 2009). In this study, Markov chain Monte Carlo estimation was used for estimation of the models, which is a Bayesian estimation technique. Bayesian estimation method requires specification of a prior distribution for each parameter in the model that reflects prior assumptions regarding that parameter. Parameter estimates may be biased if priors are poorly specified in Bayesian estimation (e.g., Mislevy, 1986). Therefore, a sufficiently informative prior should be used for obtaining accurate estimates of model parameters (Baker & Kim, 2004; Mislevy, 1986).

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). The bias in item parameter estimates due to misspecification of actual ability distribution often can be reduced by

- Selection and peer-review under responsibility of the Organizing Committee of the conference

*Corresponding author: Tugba Karadavut- E-mail: tugba-mat@hotmail.com

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).

In this study, the normality assumption for ability was investigated for its sufficiency to yield reasonable estimates of item and ability parameters, especially when the actual ability distribution was uniform. A simulation study was done to analyze student responses with a normal and a uniform underlying ability using a unidimensional IRT model for dichotomous items. These IRT models included Rasch (Rasch, 1960), two-parameter logistic (2PL; Birnbaum, 1968), and three-parameter logistic (3PL; Birnbaum, 1968) IRT models. Uniform and normal ability distributions were assumed for estimation of each IRT model. Finally, item and ability parameter estimates from these models were compared to the generating item and ability parameters. Thus, normal and uniform priors were compared for accuracy of item and ability parameter estimation in unidimensional IRT models for dichotomous items, when the actual ability distribution was normal or uniform.

Methods

Unidimensional Item Response Theory Models

The most commonly used unidimensional IRT models are the ones for dichotomous items (e.g., multiple choice). These models include Rasch, 2PL and 3PL models. The name of 2PL and 3PL models are based on the number of item parameters included in the models. Namely, the 2PL model has two item parameters, and the 3PL model has three item parameters. The 3PL model defines the probability that an examinee j with ability θ answers item i correctly ($P_i(\theta_j)$) by the following equation:

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

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 c_i parameter in 3PL model to zero reduces the 3PL model to a 2PL model. Thus, the probability that an examinee j with ability θ answers item i correctly in a 2PL model is defined as:

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

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:

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

Simulation Design

Student responses to dichotomous test items were generated using R (2016) software for Rasch, 2PL and 3PL models. Underlying ability distributions were generated 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 used to generate student responses 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

Estimation of Item Parameters

Estimation of 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:

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

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

$$\theta_j \sim \text{Normal}(0,1), \qquad j = 1, \ldots, N,$$
or (5)
$$\theta_j \sim \text{Uniform}(-4,4), \qquad j = 1, \ldots, N.$$

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 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. The scale of parameters from estimated models were placed on scale of the generating parameters by using mean and sigma equating method (Marco, 1977).

Item Recovery Analyses

Item recovery analyses were conducted to compare parameter estimation from the MCMC analyses with a normal prior and the MCMC analyses with a uniform prior. For this purpose, accuracy indices and Pearson correlations were calculated to compare item parameter estimates from the MCMC analyses with a normal prior to the generating parameters. Similarly, accuracy indices and Pearson correlations were calculates to compare item parameter estimates from MCMC analyses with a uniform prior to the generating parameters. 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 15-item and 600 sample size condition, and for 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 item difficulty (*b*) are given below:

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

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

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

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

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

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 *r*th replication, *R* is total number of replications which is 25, and *n* is the total number of items which is either 15 or 30.

Results and Findings

Accuracy indices and correlations for parameter recovery are calculated for item difficulty, item discrimination, item pseudo-guessing, and ability parameters (see Appendix A, Tables A1-A4). The MSE values did not indicate a substantial difference between uniform and normal prior for item difficulty for different IRT models. However, there were differences in MSE values for remaining item parameters and ability.

Post-hoc comparisons were conducted for transformed MSE values using Tukey's HSD procedure (see Table 2). Square-root or natural logarithm transformation was adopted for transformation of 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 uniform and normal priors for a given parameter from a particular model for a given number of items and sample size condition.

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

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

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

For 30-item and 2,000 sample size, normal prior yielded larger mean MSE values compared to uniform prior, for estimation of item difficulty and item pseudo-guessing parameter, when the actual distribution was uniform. For item discrimination, on the other hand, there was not a substantial difference in mean MSE values between normal and uniform priors. Again for 30-item and 2,000 sample size, uniform prior yielded larger mean MSE values for item difficulty, item discrimination, and for item pseudo-guessing parameter, when the actual distribution was normal. For estimation of ability in 3PL model, normal prior yielded larger mean MAE values compared to uniform prior, when the actual distribution was uniform. The effect sizes for the difference between 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 estimate for the model with normal prior, U: Mean parameter estimate 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.

Conclusion

The primary purpose of using IRT models is to locate students on a continuous scale by estimating their ability (Baker, 2001). This study compared uniform and normal distribution assumptions for estimation of item and ability parameters in IRT models. Assuming a uniform distribution for ability led to more accurate estimates of ability in the 2PL model no matter if the actual ability distribution was uniform or normal. Similarly, assuming a uniform distribution for ability in the 3PL model yielded more accurate estimates of ability when the actual ability distribution was uniform. However, the difference between assuming a normal or uniform distribution for ability was only moderate for estimation of ability, when the actual distribution was normal. For the Rasch model, there was not an explicit pattern for comparing accuracy of ability estimates from uniform and normal distribution assumptions. These results suggest using a uniform distribution assumption in a 2PL model in order to achieve more accurate estimates of ability. Similarly, a uniform distribution assumption for ability may be used for achieving more accurate estimates of ability in a 3PL model if the actual ability distribution is not known.

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.

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.

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.

Appendix A

Accuracy Indices and Correlations

Table A1: accuracy indices and correlations for 15-item and 600 sample size when actual 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

	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 15-item and 600 sample size when actual 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

	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 30-item and 2,000 sample size when actual 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

	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 30-item and 2,000 sample size when actual 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

	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