

Comparison of Parameter Recovery in a  
2-Parameter Logistic Item Response Model  
using MLE and Bayesian MCMC Methods

Tom Proctor

Khee-Shoon Teo

Jianlin Hou

Mingchuan Hsieh

The University of Iowa

## Comparison of Parameter Recovery in a 2-Parameter Logistic Item Response Model using MLE and Bayesian MCMC Methods

In psychometrics there are two major theories of test scores. The first, typically known as classical test theory (CTT) postulates that a test score is made up of a true, yet unknowable, score and the error score. Much of CTT attempts to estimate this error score which in turn gives the psychometrician estimates of such concepts as the standard error of measurement and reliability. Using CTT, test scores are usually reported as either a total correct score or in some transformation of the total correct score. Psychometricians use CTT to obtain item level statistics as well, such as item difficulty and discrimination index. Item difficulty is simply the proportion of examinees that get the item correct. The discrimination index is typically a point-biserial or biserial correlation. This index essentially indicates to what degree those who obtain a high or low total score get a particular item correct. However, nothing in CTT utilizes this information when attempting to determine an examinee's true score. For example, an examinee could get the 10 most difficult items correct and another examinee could get the 10 easiest items correct, both resulting in a total score of 10. Many people might say that such a state of affairs is unfair. This is the crux of psychometrics: how do we determine which examinees have a higher or lower level of achievement in a particular subject.

Item Response Theory (IRT) was developed in an attempt to answer that question. Perhaps the major difference between CTT and IRT is that IRT uses the information about item difficulty, discrimination, and in some models even how easy it is to guess the correct answer, as well as the mathematical model used to estimate a true score. IRT began to gain prominence as a test theory mainly in the 1950's when both Lawley and Lord separately published on IRT (du

Toit, 2003; Linn, 1989). IRT models all have two things in common, they use parameters – estimates of item characteristics – and they all attempt to estimate theta – an estimate of the underlying ability. The most popular and researched models in psychometrics are the unidimensional models. All these models have strong assumptions which include 1) unidimensionality, which requires that there be only one underlying trait for the data, and 2) local independence, which requires that when theta is held constant, the probability of examinees getting a particular item correct is independent of all other items.

The unidimensional models vary in the number of item parameters that are used. Rasch (Lord, 1980) focused on a one-parameter model, which are now known as Rasch models. This model assumed that all items have an equal discrimination index and the probability of guessing an item correctly is zero. Lord (Hambleton, 1989; Lord, 1980) focused on a three-parameter model using a normal ogive curve to represent the relationship of the probability of a correct item response versus the theta scale. In this model all three item parameters varied across items. A third model is known as the two-parameter model in which only the item difficulty and discrimination indices vary across items. Birnbaum (Lord, 1980) saw what Lord was doing and recommended using a logistic function rather than the normal ogive.

As with CTT, IRT has an item discrimination index, typically called the “a” parameter. Unlike CTT, the “a” parameter is proportional to the slope at the point on the theta scale that equals the difficulty parameter (Hambleton, 1989). The difficulty parameter, typically called the “b” parameter is the point on the theta scale at which an examinee has 50% chance of getting the item correct, this is true only for a 1 and 2 parameter model (Hambleton, 1989). Lord and Novick (du Toit, 2003) showed that when using the normal ogive both the “a” and “b” parameters are functions of the CTT discrimination index, specifically the biserial version of the index. The last

item parameter is called the guessing parameter but is really the lower asymptote parameter, and is typically called the “c” parameter. This parameter estimates what the probability of getting the item correct is when an examinee does not know the answer.

When the assumptions of IRT hold, and there is still debate about the robustness of these assumptions, the models just described offer several advantages over CTT. Specifically, if there are a large enough number of items, then the ability of an examinee can be compared to another examinee even if they did not take the same sets of items (Crocker & Algina, 1986). The converse is also true, if there is a large enough sample of examinees, the estimated item parameters can be compared to those of another item even if they were not taken by the same set of examinees (Hambleton, 1989). These advantages are what has given rise in popularity to computer adaptive testing (CAT) that is used on such tests as the Graduate Record Examination (GRE).

Certainly when high stakes decisions such as admissions to graduate school, meeting high school graduation exit exam requirements, receiving a professional license, and so on depend upon test scores it is imperative that those scores not only be precise but accurate as well. One way to determine if a particular method of estimation achieves this is to perform a simulation study. In IRT this would require simulating the item parameters and the response strings for examinees based on these parameters. This is in fact the main research question of this project. How well does the frequentist approach to estimation recover these simulated item parameters compared to using a Bayesian approach?

Perhaps the most popular software used in practice for estimating item parameters and examinee abilities is the commercially available BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). BILOG-MG uses two methods to estimate item parameters and three methods to

estimate examinee abilities. For examinee abilities the choices are maximum likelihood (ML), expected a posteriori (EAP), or maximum a posteriori (MAP). For item parameters the two methods are marginal maximum likelihood (MML) or maximum marginal a posteriori (MMAP). BILOG-MG suggests using MMAP when there are less than 250 examinees to base item parameter estimates on, or if some items are either too easy or difficult (du Toit, 2003). MMAP limits priors on the item parameters to be normal for the “b” parameter and log-normal for the “a” parameter (du Toit, 2003). The MML method uses two method to solve the marginal likelihood equations, the EM method and Newton-Gauss or Fisher scoring iterations (du Toit, 2003). To choose between MML and MMAP the user simply supplies the prior option in the item calibration step of the program.

Previous research has looked at the application of Markov Chain Monte Carlo (MCMC) methods for estimating item parameters and examinee abilities. However, several of these papers coded their own Gibbs samplers which utilizes the Metropolis-Hastings within Gibbs algorithm (Patz & Junker, 1999; Yao, Patz, & Hanson, 2002). Patz and Junker (1999) primarily focused on developing a Gibbs sampler rather than on the performance of the program to accurately estimate item parameters. None of the studies reviewed used WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2004) and all rely upon proposal densities as part of the algorithm utilized.

The Patz and Junker (1999) study used six constructed response items and 3,000 examinees from the National Assessment of Educational Progress (NAEP) and compared the results of their sampler to those from BILOG (Mislevy & Bock, 1985). This study used a two-parameter logistic (2PL) model and placed normal priors on both ability estimates ( $\theta$ ) and the difficult parameter ( $\beta$ ) (Patz & Junker, 1999). They placed a log-normal prior on the discrimination parameter ( $\alpha$ ). These priors are consistent with both BILOG-MG and NAEP

(Patz & Junker, 1999). It appears from the study that only one Markov chain was used and what the authors termed short chain – 7,400 iterations with 400 iteration burn-in – and long chain – 37,000 iterations with 2,000 iteration burn-in. Based on the short chain MCMC estimates the results compared to BILOG were essentially identical for both the item parameters and their standard errors.

Yao, Patz and Hanson (2002) utilized a three-parameter logistic (3PL) model and compared the results of a revised Gibbs sampler from the Patz and Junker study to estimates obtained by PARDUX (Burket, 1996) which also estimates IRT item and ability parameters using MML and ML. This study used a normal prior on beta, a log-normal on alpha, and beta on the guessing parameter. Aside from the choice of IRT model the major difference between this study and the previous study is that this study used both real and simulated data.

The real test data consisted of 15 items, of which 10 were multiple choice and 5 were constructed response item, and 1,200 examinees (Yao, Patz, & Hanson, 2002). Again it appears from the study that only one Markov chain was used and in this case 6,800 iterations with a 700 iteration burn-in. The results on the actual test data sample resulted in correlations of the three parameters based on MCMC and PARDUX to be 0.975, 0.997, and 0.998. The parameter estimates were similar in magnitude as well.

The simulated test data consisted of 16 items, 10 of which were dichotomous items and six were polytomous items. No explicit mention is made on how the data for this part of the study was simulated and one is left assuming that it was based on the priors mentioned earlier. Two thousand examinee responses were simulated using a normal (0, 1) distribution. Again, it appears only one Markov chain was used and this time with 6,000 iterations with a 300 iteration burn-in. The authors offer no conclusions or discussions of how either MCMC or PARDUX

recovered the true item parameters based on simulation. A review to the table provided by the study makes it difficult to say that one or the other method recovered any or all three parameters better.

Jones and Nediak (2000) looked at actual test data for 28,184 examinees and also 21,116 simulated examinees for 101 items on the Law School Admissions Test (LSAT). This study also used a modification to the Patz and Junker (1999) sampler. Results of the MCMC were compared to estimates from BILOG utilizing a 3PL model. As with the previous two studies it appears only one Markov chain was used and in this case 7,000 iterations with a 2,000 iteration burn-in was used. The priors on the item parameters were uniform on a rectangular region of  $(-6, 6)$  by  $(.5, 2.5)$  by  $(0.0, 0.5)$ .

As with the previous studies when applied to actual test data the estimates between BILOG and MCMC were similar. When applied to simulated data with known parameters the MCMC method appears to recover the known parameters better than BILOG. In most cases MCMC estimates differed from the known parameter by 0.05, whereas BILOG differed by over  $1/10^{\text{th}}$ .

The cursory review of the literature suggests that the use of Bayesian methods and MCMC can lead users of IRT to better estimates of item parameters. All three studies reviewed state that the use of MCMC results in estimates of the joint posterior distribution of the item parameters whereas programs such as BILOG and PARDUX give only point estimates (Jones & Nediak, 2000; Patz & Junker, 1999; Yao, Patz, & Hanson, 2002).

A major drawback to typical users of IRT is the accessibility of the MCMC methods. The studies reviewed here either coded the sampler in S-Plus or developed proprietary software. WinBUGS is a free downloadable program that uses Bayesian and MCMC methods that is

available to any user of IRT. This project will make use of WinBUGS to see how well it recovers known item parameters versus BILOG-MG utilizing a 2PL model:

$$p_i[\theta] = \frac{1}{1 + e^{-1.7\alpha_i(\theta - \beta_i)}} \quad (1)$$

where  $\alpha$  is the discrimination parameter,  $\beta$  is the difficulty parameter and  $\theta$  is the ability parameter. The constant of 1.7 puts the logistic function on a “normal” metric.

### Methods

The data for this project was simulated using a SAS routine. Parameters for 30 items were simulated using a log-normal (0, 4)<sup>1</sup> for alpha and a normal (0, 1) for beta. For theta a normal (0, 1) was used to simulate 1,000 examinees. Based on these simulated parameters, 1,000 item response strings, coded 0 for wrong and 1 for correct, was obtained. This was done 15 times and 15 different sets of item parameters, thetas, and item response datasets were generated. The dataset with the fewest number of simulated examinees who either got all items correct or all items incorrect was chosen. The reason for doing this was to minimize problems that ML has in estimating theta for perfect scores. From the 1,000 response strings, a random sample of 500 of these response strings was obtained as a second sample.

The simulated data was run through BILOG-MG to obtain item parameters estimates based on MML. The same data was then run through WinBUGS using two different sets of priors on the item parameters. The first run, termed the exact prior, placed a normal (0, 1) prior on beta and a log-normal (0, 4) alpha. The second run, termed the non-informative prior, place a normal (0, 0.001) prior on beta and a log-normal (0, 0.001) prior on alpha. For both runs of WinBUGS the prior on theta remained normal (0, 1). Three separate chains were run for each set

---

<sup>1</sup> From this point forward the parameterization of normal and log-normal distributions is in terms of mean and precision which is consistent with WinBUGS.

of priors. Initial starting values were generated using WinBUGS. See Appendix A for WinBUGS code.

### Results

For the sample of 1,000 simulated examinees 500 iterations were used by WinBUGS in its adaptation phase, an additional 500 iterations were used for additional burn-in and an additional 1,000 iterations were used to estimate to posterior means of alpha and beta. For the sample size of 500 simulated examinees, 500 iterations were used by WinBUGS in its adaptation phase, an additional 1,000 iterations for the exact priors and 2,500 iterations for the non-informative priors were used for burn-in, and to estimate the posterior means for an additional 2,500 iterations were for the exact prior and 1,500 for the non-informative prior. Below are representative history, BGR, and autocorrelation plots for alpha and beta based on all possible iterations based on the sample size of 1,000. The autocorrelation and BGR statistics suggest that chains reached an acceptable convergence on the stationary distribution. The plots are also similar for the sample size of 500.

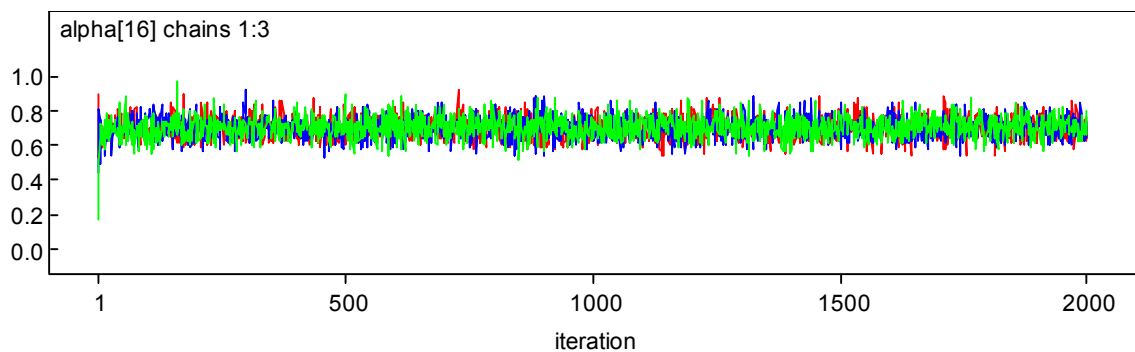


Figure 1a. MCMC history plot for alpha on Item 16

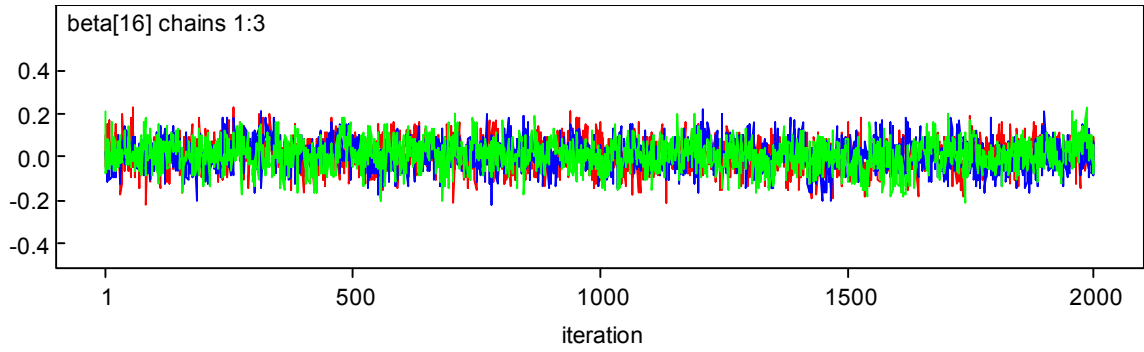


Figure 1b. MCMC history plot for beta on Item 16

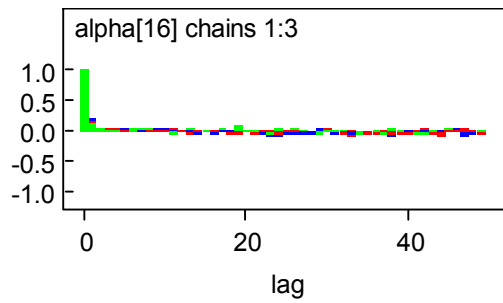


Figure 1c. Autocorrelation plot for alpha on Item 16

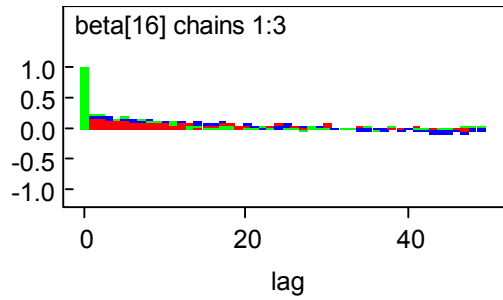


Figure 1d. Autocorrelation plot for beta on Item 16

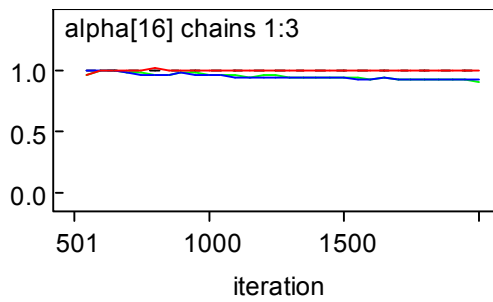


Figure 1e. BGR statistic for alpha on Item 16

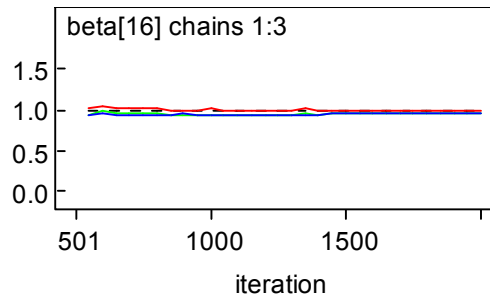


Figure 1f. BGR statistic for beta on Item 16

For the sample size of 1,000 Table 1a lists estimates of alpha and the “true” alpha for the 30 simulated items. Table 1b lists estimates of beta and the “true” beta for the items. Figures 2a and 2b provide a graphical representation of the parameter estimates for all three methods and true values.

Estimates based on the ML method using BILOG-MG resulted in the true alpha parameter not being included in a 95% confidence interval for items 1, 12, 18 and 23. As for the beta parameter estimate the ML method resulted in the true parameter not being included in a 95% confidence interval for items 15, 16, 18, and 22. Using exact priors in WinBUGS resulted in the true alpha parameter not being included in a 95% credible set for items 1 and 18, and for the true beta parameter for items 22 and 24 not being included in the 95% credible set. For the non-informative prior in WinBUGS the true alpha parameter was not covered by the 95% credible set for items 1, 18, and 23, and the true beta parameter for items 1 and 16 were not covered by a 95% credible set.

Table 1a. Estimates of the alpha parameter, N = 1,000

ITEM	MML		Exact prior		Non-informative prior		True $\alpha$
	$\tilde{\alpha}$	SD	$\tilde{\alpha}$	SD	$\tilde{\alpha}$	SD	
1	0.33	0.05*	0.36	0.04*	0.31	0.08*	0.45
2	0.88	0.08	0.90	0.08	0.88	0.06	0.95
3	0.66	0.06	0.68	0.06	0.65	0.21	0.69
4	2.26	0.20	2.29	0.20	2.29	0.08	2.37
5	1.05	0.08	1.07	0.08	1.05	0.14	1.09
6	1.37	0.15	1.41	0.14	1.37	0.07	1.36
7	0.87	0.07	0.89	0.07	0.87	0.08	0.84
8	0.92	0.07	0.94	0.08	0.91	0.12	0.87
9	1.32	0.13	1.36	0.12	1.32	0.06	1.32
10	0.66	0.06	0.69	0.07	0.66	0.16	0.67
11	1.47	0.16	1.50	0.16	1.47	0.15	1.46
12	1.60	0.14*	1.63	0.15	1.60	0.15	1.88
13	1.32	0.14	1.36	0.15	1.32	0.09	1.25
14	0.87	0.08	0.91	0.09	0.87	0.12	0.89
15	1.49	0.11	1.52	0.13	1.50	0.06	1.40
16	0.68	0.06	0.70	0.06	0.68	0.13	0.71
17	1.63	0.13	1.67	0.14	1.64	0.11	1.61
18	1.40	0.11*	1.43	0.11*	1.40	0.07*	1.12
19	0.70	0.06	0.72	0.06	0.69	0.08	0.8
20	0.84	0.08	0.88	0.08	0.84	0.06	0.91
21	0.67	0.06	0.69	0.06	0.67	0.08	0.70
22	0.98	0.08	1.01	0.08	0.98	0.04	0.91
23	0.32	0.04*	0.35	0.04	0.31	0.08*	0.41
24	0.84	0.07	0.86	0.08	0.83	0.05	0.77
25	0.52	0.05	0.54	0.05	0.51	0.07	0.54
26	0.52	0.06	0.57	0.06	0.50	0.10	0.55
27	1.30	0.09	1.33	0.11	1.31	0.08	1.44
28	0.61	0.08	0.65	0.07	0.60	0.08	0.63
29	0.94	0.08	0.97	0.08	0.94	0.08	1.01
30	0.74	0.08	0.79	0.08	0.74	0.08	0.85

Note: An “\*” indicates that the true parameter is not within the 95% confidence interval for the

MLE or a 95% credible set for the MCMC methods.

Table 1b. Estimates of the beta parameter, N = 1,000

ITEM	MML		Exact prior		Non-informative prior		True $\beta$
	$\tilde{b}$	SD	$\tilde{b}$	SD	$\tilde{b}$	SD	
1	1.32	0.21	1.20	0.17	1.40	0.24*	1.02
2	-1.34	0.10	-1.32	0.10	-1.35	0.10	-1.26
3	-0.04	0.06	-0.05	0.07	-0.04	0.07	-0.01
4	0.16	0.03	0.14	0.04	0.16	0.04	0.13
5	-0.36	0.05	0.37	0.06	-0.37	0.06	-0.39
6	1.56	0.09	1.51	0.09	1.56	0.09	1.43
7	-0.46	0.06	0.46	0.06	-0.46	0.07	-0.51
8	0.88	0.06	0.82	0.07	0.86	0.08	0.81
9	-1.16	0.06	-1.14	0.07	-1.17	0.07	-1.22
10	-1.52	0.13	-1.49	0.13	-1.54	0.14	-1.55
11	1.71	0.09	1.66	0.10	1.71	0.10	1.61
12	1.12	0.05	1.08	0.07	1.12	0.06	1.12
13	1.91	0.11	1.85	0.11	1.92	0.12	1.89
14	1.63	0.11	1.57	0.12	1.65	0.13	1.58
15	-0.59	0.04*	-0.59	0.05	-0.59	0.05	-0.68
16	0.02	0.06*	0.00	0.07	0.02	0.07*	-0.12
17	-0.69	0.04	-0.69	0.05	-0.70	0.05	-0.64
18	0.64	0.04*	0.60	0.05	0.62	0.05	0.53
19	1.23	0.10	1.18	0.10	1.24	0.11	1.11
20	-1.65	0.12	-1.60	0.12	-1.67	0.13	-1.48
21	-0.65	0.08	-0.64	0.08	-0.66	0.08	-0.74
22	-0.72	0.06*	-0.71	0.06*	-0.72	0.07	-0.85
23	0.46	0.13	0.41	0.13	0.48	0.15	0.5
24	1.28	0.09	1.23	0.10*	1.29	0.10	1.44
25	-0.87	0.11	-0.85	0.11	-0.88	0.11	-0.76
26	2.30	0.25	2.11	0.21	2.41	0.29	2.2
27	0.29	0.04	0.27	0.05	0.29	0.05	0.26
28	-2.33	0.24	-2.21	0.20	-2.39	0.27	-2.49
29	1.14	0.08	1.10	0.08	1.15	0.09	1.14
30	1.92	0.16	1.83	0.15	1.95	0.17	1.74

Note: An “\*” indicates that the true parameter is not within the 95% confidence interval for the

MLE or a 95% credible set for the MCMC methods.

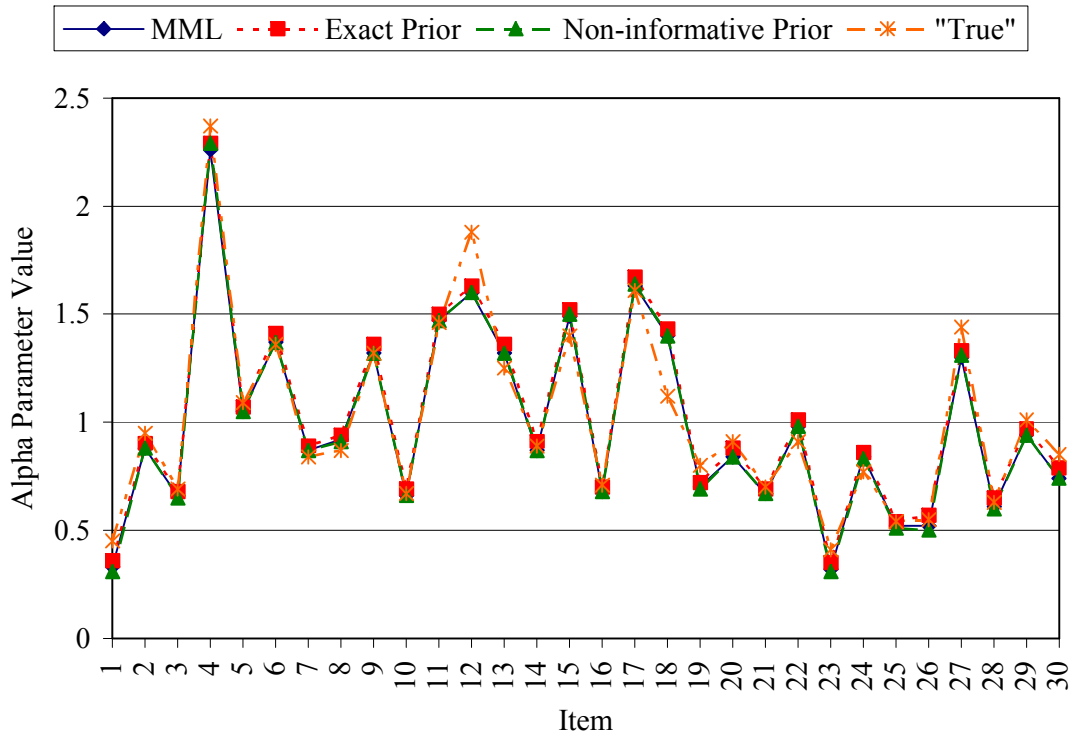


Figure 2a. Comparison of Alpha Parameters, N = 1,000

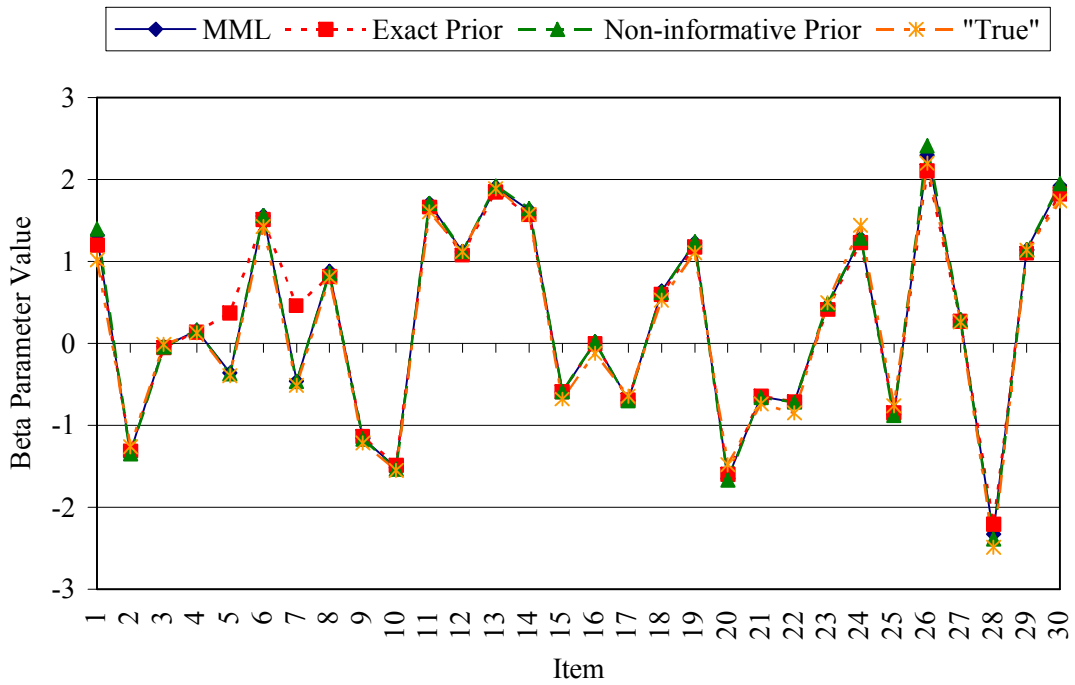


Figure 2b. Comparison of Beta Parameters, N = 1,000

For the sample size of 500 Table 2a lists estimates of alpha and the “true” alpha for the 30 simulated items. Table 2b lists estimates of beta and the “true” beta for the items. Figures 2a and 2b provide a graphical representation of the parameter estimates for all three methods and true values.

Estimates based on the ML method using BILOG-MG resulted in the true alpha parameter not being included in a 95% confidence interval for items 12, 18, 19, 23 and 27. As for the beta parameter estimate the ML method resulted in the true parameter being included in the 95% confidence interval for all items. Using exact priors in WinBUGS resulted in the true alpha parameter not being included in a 95% credible set for item 18, and for the true beta parameter was included in the 95% credible set for all the items. For the non-informative prior in WinBUGS the true alpha parameter was not covered by the 95% credible set for items 18, 19, 23, 27, and 30, and the true beta parameter for items 25 and 30 were not covered by a 95% credible set.

Table 2a. Estimates of the alpha parameter, N = 500

ITEM	MML		Exact Prior		Non-informative Prior		True $\alpha$
	$\tilde{a}$	SD	$\tilde{a}$	SD	$\tilde{a}$	SD	
1	0.36	0.06	0.40	0.06	0.33	0.07	0.45
2	0.84	0.12	0.88	0.11	0.81	0.11	0.95
3	0.63	0.08	0.66	0.08	0.61	0.08	0.69
4	2.15	0.28	2.12	0.26	2.13	0.28	2.37
5	0.99	0.11	1.03	0.11	0.98	0.11	1.09
6	1.39	0.21	1.45	0.21	1.41	0.21	1.36
7	0.89	0.09	0.92	0.10	0.87	0.10	0.84
8	0.91	0.10	0.94	0.11	0.89	0.10	0.87
9	1.14	0.16	1.18	0.15	1.11	0.15	1.32
10	0.67	0.09	0.72	0.09	0.65	0.09	0.67
11	1.39	0.18	1.44	0.21	1.38	0.21	1.46
12	1.47*	0.18	1.53	0.19	1.49	0.19	1.88
13	1.30	0.20	1.37	0.21	1.31	0.22	1.25
14	1.02	0.13	1.07	0.13	1.00	0.14	0.89
15	1.35	0.13	1.39	0.16	1.34	0.16	1.40
16	0.72	0.08	0.75	0.09	0.70	0.08	0.71
17	1.38	0.15	1.43	0.16	1.37	0.16	1.61
18	1.45*	0.16	1.49*	0.17	1.44*	0.17	1.12
19	0.63*	0.08	0.67	0.08	0.60*	0.09	0.80
20	0.83	0.11	0.87	0.12	0.79	0.12	0.91
21	0.59	0.08	0.62	0.08	0.57	0.08	0.70
22	1.07	0.13	1.12	0.13	1.06	0.13	0.91
23	0.27*	0.06	0.32	0.06	0.24*	0.07	0.41
24	0.92	0.11	0.97	0.12	0.91	0.11	0.77
25	0.54	0.07	0.58	0.08	0.52	0.08	0.54
26	0.64	0.10	0.70	0.10	0.61	0.10	0.55
27	1.19*	0.12	1.22	0.13	1.17*	0.12	1.44
28	0.70	0.13	0.76	0.11	0.65	0.12	0.63
29	0.88	0.11	0.92	0.11	0.86	0.11	1.01
30	0.65	0.11	0.72	0.10	0.61*	0.11	0.85

Note: An “\*” indicates that the true parameter is not within the 95% confidence interval for the

MLE or a 95% credible set for the MCMC methods.

Table 2b. Estimates of the beta parameter, N = 500

ITEM	MML		Exact Prior		Non-informative Prior		True $\beta$
	$\tilde{b}$	SD	$\tilde{b}$	SD	$\tilde{b}$	SD	
1	1.08	0.23	0.96	0.20	1.21	0.31	1.02
2	-1.34	0.13	-1.30	0.14	-1.40	0.15	-1.26
3	-0.04	0.09	-0.05	0.10	-0.04	0.10	-0.01
4	0.15	0.04	0.13	0.06	0.14	0.06	0.13
5	-0.45	0.07	-0.45	0.08	-0.46	0.09	-0.39
6	1.62	0.12	1.55	0.13	1.63	0.13	1.43
7	-0.43	0.08	-0.42	0.09	-0.44	0.09	-0.51
8	0.85	0.09	0.81	0.10	0.87	0.11	0.81
9	-1.33	0.11	-1.29	0.11	-1.36	0.13	-1.22
10	-1.56	0.18	-1.49	0.17	-1.63	0.21	-1.55
11	1.71	0.12	1.64	0.13	1.74	0.15	1.61
12	1.13	0.08	1.08	0.09	1.15	0.10	1.12
13	1.91	0.15	1.82	0.15	1.95	0.18	1.89
14	1.51	0.13	1.44	0.13	1.55	0.15	1.58
15	-0.69	0.06	-0.67	0.08	-0.69	0.08	-0.68
16	-0.02	0.08	-0.03	0.09	-0.02	0.10	-0.12
17	-0.75	0.07	-0.73	0.08	-0.75	0.08	-0.64
18	0.57	0.06	0.54	0.07	0.58	0.07	0.53
19	1.27	0.16	1.19	0.15	1.33	0.19	1.11
20	-1.68	0.16	-1.62	0.17	-1.76	0.20	-1.48
21	-0.84	0.13	-0.81	0.13	-0.88	0.15	-0.74
22	-0.88	0.08	-0.85	0.09	-0.89	0.10	-0.85
23	0.38	0.22	0.31	0.19	0.60	2.26	0.50
24	1.34	0.12	1.27	0.13	1.37	0.14	1.44
25	-1.03	0.16	-0.98	0.15	-1.09*	0.18	-0.76
26	2.02	0.25	1.85	0.22	2.14	0.31	2.20
27	0.29	0.06	0.27	0.07	0.30	0.08	0.26
28	-2.26	0.31	-2.13	0.24	-2.47	0.39	-2.49
29	1.15	0.11	1.10	0.12	1.18	0.13	1.14
30	2.17	0.28	1.99	0.23	2.34*	0.36	1.74

Note: An “\*” indicates that the true parameter is not within the 95% confidence interval for the MLE or a 95% credible set for the MCMC methods.

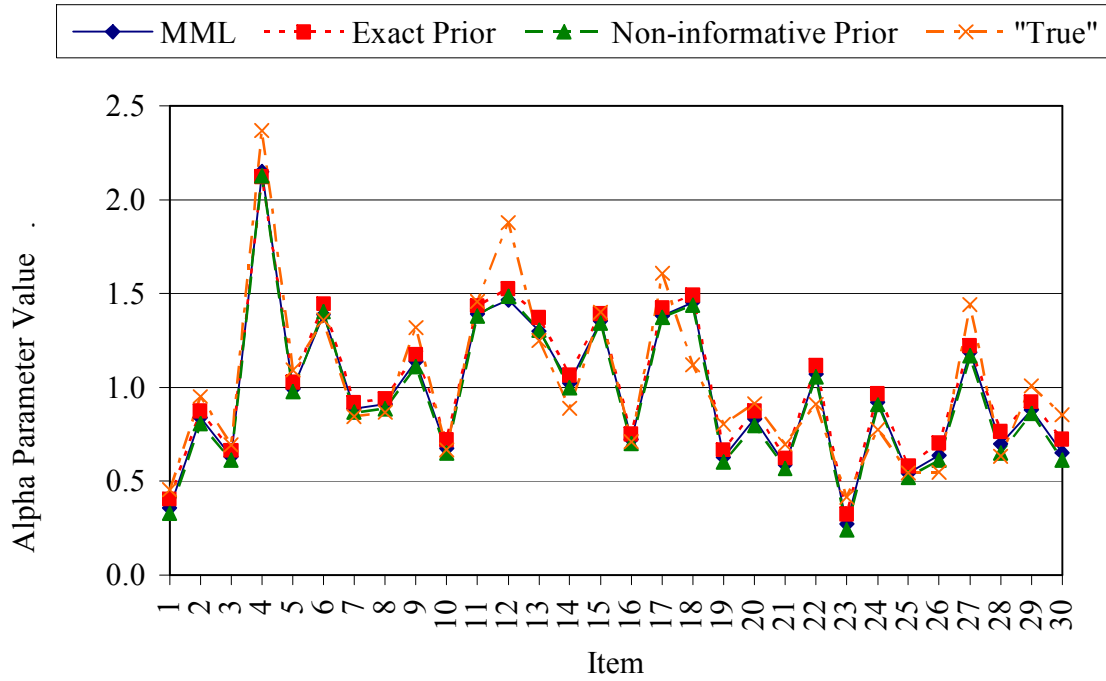


Figure 3a. Comparison of Alpha Parameters, N = 500

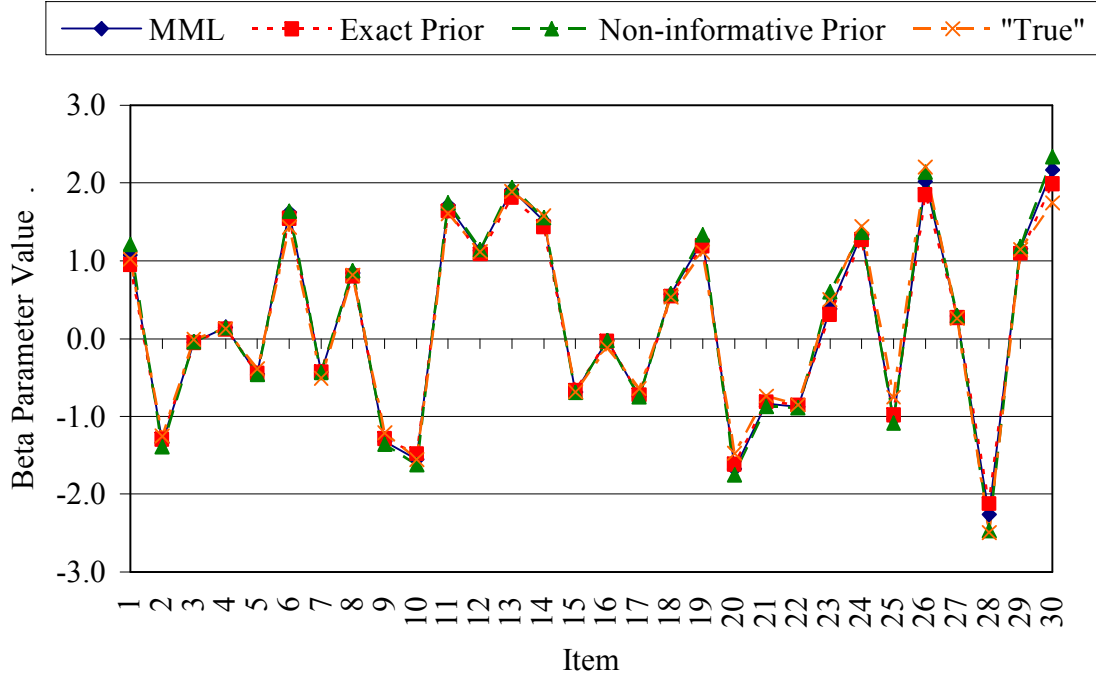


Figure 3b. Comparison of Beta Parameters, N = 500

## Discussion

Overall, the use of WinBUGS to estimate item parameters for a 2PL IRT model resulted in better recovery of simulated item parameters compared to those through the use of BILOG-MG when using a sample size of 1,000. When a sample size of 500 was used the results are mixed. Using what has been termed the exact prior, those based on the parameters of the distributions used to create the simulated data, only slightly outperformed using a non-informative prior in the absolute number of true item parameters recovered, again for the sample size of 1,000. For the smaller sample size, the use of the “exact” prior recovered parameters better than the use of non-informative priors. With regards to the alpha parameter, and the sample size of 1,000, the exact prior failed to recover only two items, determined by whether or not the true parameter was within the 95% credible set. The non-informative prior failed to recover three items, two of which were the same as when the exact prior was used. The MML method failed to recover four items, three of which were the same as when the non-informative prior was used. All three methods failed to recover the true alpha parameter for items 1 and 18. When the sample size was 500 the exact prior only failed to recover the alpha parameter for item 18. Both the non-informative prior and the MML methods failed to recover five items, items 18, 19, 23, and 27 were common, and in addition MML failed to recover item 12 and the non-informative prior failed to recover item 30 in addition. As with the larger sample size all three methods failed to recover the alpha parameter for item 18. However, this time all three methods did recover the alpha parameter for item 1. In the case of item 1, the estimates for smaller sample size all increase and got closer to true parameter value while the standard error of the estimates remain comparable for this item across sample sizes. In the case of item 18, for the smaller

sample size the item parameters again increased but in this case moved further from the actual parameter value and the standard error of the estimates increased comparably.

For the sample size of 1,000 the results reveal a similar picture with regards to the beta parameter. The use of an exact prior and a non-informative prior resulted in two items not being recovered. However, there was no overlap in the items not recovered by these prior assumptions. The MML again failed to recover four items. Two of these items overlap with one item that each of the exact and non-informative prior failed to recover. MML failed to recover both parameters for item 18 and the non-informative prior failed to recover both parameters for item 1. All three methods failed to recover at least one parameter on each of these items suggesting that neither of these items may fit the data well. It is also worth noting that the MML and MCMC method resulted in exactly the same parameter estimate but the true parameter was not in corresponding confidence interval but was include in the credible set, for example the alpha parameter for item 12. When the sample size is reduced to 500 the results for the beta parameter is quite different than for the larger sample. Both the MML and exact prior methods recovered the beta parameter for all the items. The non-informative prior again failed to recover the beta parameter for two items; however these are different items than the ones not recovered by the larger sample size for this prior assumption.

At first sight it appears that using the exact priors and a smaller sample size would be ideal. First, the smaller sample size was randomly sampled from the larger sample and may have capitalized on chance. If accurate and precise estimates of the parameters could be obtained using a smaller sample this would be ideal for testing companies and the costs of obtaining this data. However, the use of exact prior stacked the odds in favor of WinBUGS to perform better than BILOG-MG or even WinBUGS with a non-informative prior. More research should be

done using different priors and simulated samples of differing sizes before any conclusions about MCMC outperforming BILOG-MG based on sample size. Also, with the smaller sample size used in this project, most of the standard errors increased compared to the larger sample size, making it more likely the smaller sample size confidence intervals or credible sets would cover the true parameter value. Research should also look at what happens to estimates of the latent ability and ultimately the scores reported to examinees and users of the scores. The consequential validity of test scores should out-weigh the costs associated with obtaining both accurate and precise item parameter estimates.

Based on the results of this project, using WinBUGS even with a non-informative prior resulted in better recovery of item parameters over BILOG-MG when using a sample size of 1,000. This of course requires that one accept the priors used. One option within BILOG-MG is to use MMAP which assumes the same prior distributions that were used in this project so it seems reasonable to accept the priors as ones typically used in psychometrics. One question of interest is the prior values of the parameters of these prior distributions. It may be safe to assume that if MCMC recovers item parameters better than MML on simulated data it would perform similarly on actual test data. This should result in better estimates of an examinee's ability and ultimately in better decisions based on test scores.

Our results have shown that the parameters of certain items were not recovered. This could either be due to the "luck of the draw" or something inherent in the estimation approaches. We hypothesize that it is the former rather than the latter. This is because the true parameter values for these "problematic" items were not even outliers. If other item parameters of similar true values could be recovered, there is no reason why these could not, except for the randomness of the coin-flip, so to speak. One possible extension of this project would be in

analyzing the reliabilities of the estimation approaches. This could be done by generating more datasets of item response strings while using the same set of 30 item parameters and 1000 examinees thetas, and repeating the previous analyses.

## References

- Burket, G. R. (1996). PARDUX (Version 4.1). Monterey, CA: CTB/McGraw-Hill.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
- du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International, Inc.
- Hambleton, R. K. (1989). Principles and Selected Applications of Item Response Theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Phoenix, AZ: Oryx Press.
- Jones, D. H., & Nediak, M. (2000). *Item parameter calibration of LSAT items using MCMC approximation of Bayes posterior distributions*. (No. RRR 7-2000). Piscataway, NJ: RUTCOR.
- Linn, R. L. (1989). Current Perspectives and Future Directions. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Phoenix, AZ: The Oryx Press.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: L. Erlbaum Associates.
- Mislevy, R., & Bock, D. (1985). BILOG. Chicago: Scientific Software International, Inc.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146-178.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2004). WinBUGS (Version 1.4.1). Cambridge: Imperial College.

Yao, L., Patz, R. J., & Hanson, B. A. (2002). More efficient Markov Chain Monte Carlo estimation in IRT using marginal posteriors. Retrieved Oct. 29, 2005, from

<http://www.ncme.org/repository/incoming/86.pdf#search='yao%20patz%20irt'>

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (1996). BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items (Version 3.0). Chicago: Scientific Software International, Inc.

## Appendix A

## WinBUGS Code

```

#2PL Model
model {
  for (k in 1:I) {
    alpha[k] ~ dlnorm(mu1,pre1)
    beta[k] ~ dnorm(mu2,pre2)
  }
  for (j in 1:N) {
    for (k in 1:I) {
      p[j,k] <- (exp(1.7*alpha[k]*(theta[j]-beta[k]))) / (1+(exp(1.7*alpha[k]*(theta[j]-beta[k]))))
      # logit(p[j,k]) <- 1.7*alpha[k]*(theta[j]-beta[k])
      # p[j,k] <- (1)/(1+(exp(-1.7*alpha[k]*(theta[j]-beta[k]))))
      r[j,k] ~ dbern(p[j,k])
    }
    theta[j] ~ dnorm(0,1)
  }
  # Exact Prior
  mu1 <- 0
  pre1 <- 4
  mu2 <- 0
  pre2 <- 1
  #Non-informative Prior
  # mu1 <- 0
  # pre1 <- 0.001
  #mu2 <- 0
  #pre2 <- 0.001
}

```