A Comparison of Item Parameter Estimates from Xcalibre 4.1 and Bilog-MG

David J. Weiss and Shannon Von Minden

Technical Report

February 2012

Suite 200, 2233 University Avenue
St. Paul MN 55114, U.S.A.
www.assess.com

A Comparison of Item Parameter Estimates from Xcalibre 4.1 and Bilog-MG

The purpose of this monte-carlo simulation study was to compare the item parameter estimates from Xcalibre (version 4.1 beta; Guyer & Thompson, 2011a) and Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Both programs use marginal maximum likelihood estimation (Bock & Aitkin, 1981) to estimate item parameters. The goal was to compare how well each program's estimates recovered the true item parameters and also to investigate the relationships between estimates from the two programs. The estimation programs were compared under the two-parameter logistic model (2PLM) and the three-parameter logistic model (3PLM).

Method

There were three distribution conditions for the true item parameters. True item parameters were generated from uniform, normal, or positively skewed distributions. In the uniform distribution condition, the discrimination (*a)* parameters ranged from 0.25 to 1.75, the difficulty (*b*) parameters ranged from −3 to 3, and the guessing (*c*) parameters ranged from 0.20 to 0.30. In the normal distribution condition, the *a* parameters had a mean of 1 and a standard deviation (SD) of 0.25, the *b* parameters had a mean of 0 and an SD of 1, and the *c* parameters had a mean of 0.25 and an SD of 0.02. In the skewed condition, the same location and scale parameters from the normal distribution condition were used, but all of the item parameters were drawn from distributions that had skewness values of 2.

Two test lengths were used: short tests had 25 items and long tests had 50 items. There were also two sample sizes: small samples had $N = 200$ examinees and large samples had $N =$ 1,000 examinees. All of the examinee θ values were generated from a normal distribution with a mean of 0 and an SD of 1. Item responses were simulated using a program written in R (R Development Core Team (2010). Probabilities of correct responses for an item were calculated from the three-parameter logistic model (3PLM),

$$P_{ij}\left(\theta_j\right) = c_i + \left(1 - c_i\right)\frac{\exp\left[\mathrm{D}a_i\left(\theta_j - b_i\right)\right]}{1 + \exp\left[\mathrm{D}a_i\left(\theta_j - b_i\right)\right]} \qquad (1)$$

where θ_j is the trait level for examinee j; a_i is the discrimination, b_i is the difficulty, and c_i is the psuedoguessing parameter for item i; and D = 1.70. For the two-parameter model (2PLM), c was fixed at 0.0. For each item-by-simulee interaction, a random number between 0 and 1 was generated from a uniform distribution. If the random number was less than or equal to the probability of a correct response, the item was scored as correct (1); otherwise it was scored 0.

Prior distributions are often used when estimating item parameters, and can be particularly useful for short tests or when the test was administered to a small number of examinees, because prior distributions help to ensure finite parameter estimates that are within a particular range. Default priors were used in Bilog because many users are likely to select the default priors when estimating item parameters. In Bilog, the default prior for the log *a* parameters was normal with a mean of 0 and an SD of 0.50, the default prior for the *b* parameters

was normal with a mean of 0 and an SD of 2, and the default prior for the *c* parameters was beta with shape parameters $20p + 1$ and $20(1 - p) + 1$, where p = 1/number of alternatives.

In Xcalibre, the default prior for the *a* parameters was normal with a mean of 0.80 and an SD of 0.20, the default prior for the *b* parameters was normal with a mean of 0 and an SD of 1, and the default prior for the *c* parameters was normal with a mean of *p* and an SD of 0.03. When using Xcalibre, users can use either traditional priors or choose the option of "floating priors," in which case the mean of the item parameters after a given loop is used as the new (updated) prior distribution mean.

To compare item parameter estimates for the 3PLM, item parameters were estimated using no priors (Bilog) or floating priors (Xcalibre), and in another condition fixed (i.e., not floating) priors were used in both programs.

Results

2PLM: No Priors/Floating Priors

The means and SDs of the true and estimated *a* and *b* parameters are shown in Tables 1 and 2, respectively. The mean item discriminations for both programs were very close to the corresponding true means; Bilog means tended to be higher than the true means and Xcalibre means tended to be slightly lower. With the exception of the N = 1,000 condition with 50 items, Xcalibre means were closer to the true means for skewed true distributions. The SDs of the estimated item discriminations using Xcalibre were around 0.20 (the SD of the prior) regardless of the true values. Consequently, the SD of estimated item discriminations were also closer to the true values for Xcalibre.

Table 1. Means and Standard Deviations of Item Discriminations for Uniform, Normal, and Skewed True Parameter Distributions

Condition and	True		Xcalibre		Bilog	
Distribution	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	0.871	0.476	0.802	0.254	0.871	0.500
Normal	0.976	0.287	0.926	0.226	1.003	0.305
Skewed	1.166	0.197	1.112	0.259	1.185	0.314
N = 200, 50 Items						
Uniform	1.071	0.424	0.930	0.242	1.073	0.478
Normal	0.907	0.240	0.848	0.219	0.912	0.323
Skewed	1.172	0.132	1.188	0.204	1.202	0.248
N = 1,000, 25 Items						
Uniform	0.871	0.476	0.785	0.266	0.915	0.465
Normal	0.976	0.287	0.878	0.221	1.000	0.323
Skewed	1.166	0.197	1.248	0.321	1.256	0.261
N = 1,000, 50 Items						
Uniform	1.071	0.424	0.978	0.339	1.135	0.552
Normal	0.907	0.240	0.866	0.185	0.932	0.234
Skewed	1.172	0.132	1.23	0.223	1.209	0.185

The mean estimated item difficulties (Table 2) were also all very close to the true values, with Xcalibre performing slightly better than Bilog for eight of twelve conditions, including all skewed true distributions. The SDs of the estimated item difficulties were very similar for the two programs, and all were close to the true values with no clear trends for either program.

Table 2. Means and Standard Deviations of Item Difficulties for Uniform, Normal, and Skewed True Parameter Distributions

Condition and	True		Xcalibre		Bilog	
Distribution	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	0.053	1.695	0.005	1.721	−0.100	1.793
Normal	−0.176	0.996	−0.226	1.073	−0.245	1.031
Skewed	0.707	0.666	0.695	0.695	0.662	0.666
N = 200, 50 Items						
Uniform	0.126	1.904	0.098	1.889	0.081	1.935
Normal	0.283	1.128	0.266	1.140	0.280	1.144
Skewed	0.582	0.612	0.548	0.625	0.532	0.615
N = 1,000, 25 items						
Uniform	0.053	1.695	0.065	1.826	0.050	1.650
Normal	−0.176	0.996	−0.218	1.053	−0.196	0.959
Skewed	0.707	0.666	0.671	0.663	0.649	0.635
N = 1,000, 50 Items						
Uniform	0.126	1.904	0.044	1.917	0.040	1.881
Normal	0.283	1.128	0.239	1.119	0.230	1.083
Skewed	0.582	0.612	0.558	0.621	0.556	0.609

The RMSE values of the item parameter estimates are shown in Table 3. For the *a* parameters, most of the RMSE values were around 0.10 or 0.20 for both programs. The exception was the 50-item test with $N = 200$ and a uniform parameter distribution. In that condition, the RMSE values for both programs were close to 0.30. The RMSE values for the *b* parameters were uniformly smaller for Xcalibre with $N = 200$ regardless of test length. In general, RMSE was largest when the item parameters were uniformly distributed, smaller when the parameters were normally distributed, and smallest when the parameters were skewed. RMSE values were also smaller when a large sample size was used.

The bias values of the item parameter estimates are shown in Table 4. When using Xcalibre, both *a* and *b* parameters were generally underestimated. When using Bilog however, the *a* parameters were typically overestimated while the *b* parameters were underestimated. Xcalibre generally had slightly more biased estimates of the *a* parameters. For the *b* parameters, however, Xcalibre generally had less biased estimates than did Bilog.

Table 3. RMSE of Item Parameter Estimates for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	Xcalibre *a*	Xcalibre *b*	Bilog *a*	Bilog *b*
N = 200, 25 Items				
Uniform	0.246	0.276	0.148	0.374
Normal	0.227	0.209	0.227	0.245
Skewed	0.229	0.078	0.267	0.115
N = 200, 50 Items				
Uniform	0.305	0.330	0.335	0.410
Normal	0.156	0.162	0.188	0.198
Skewed	0.157	0.125	0.185	0.131
N = 1,000, 25 Items				
Uniform	0.253	0.226	0.175	0.172
Normal	0.144	0.106	0.077	0.094
Skewed	0.209	0.077	0.148	0.087
N = 1,000, 50 Items				
Uniform	0.221	0.224	0.305	0.222
Normal	0.101	0.092	0.082	0.105
Skewed	0.188	0.070	0.120	0.065

Table 4. Bias of Item Parameter Estimates for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	Xcalibre *a*	Xcalibre *b*	Bilog *a*	Bilog *b*
N = 200, 25 Items				
Uniform	−0.069	−0.048	0.000	−0.152
Normal	−0.050	−0.050	0.027	−0.069
Skewed	−0.055	−0.012	0.019	−0.046
N = 200, 50 Items				
Uniform	−0.141	−0.029	0.002	−0.045
Normal	−0.058	−0.017	0.005	−0.003
Skewed	0.016	−0.034	0.029	−0.050
N = 1,000, 25 Items				
Uniform	−0.086	0.012	0.043	−0.002
Normal	−0.098	−0.042	0.024	−0.020
Skewed	0.082	−0.036	0.089	−0.058
N = 1,000, 50 Items				
Uniform	−0.093	−0.082	0.065	−0.086
Normal	−0.041	−0.044	0.025	−0.053
Skewed	0.062	−0.024	0.035	−0.026

The correlations between item parameter estimates from the two programs are shown in Table 5. All of the estimates were highly related. In general, the *b* parameters had slightly higher correlations than the *a* parameters. Item difficulties correlated .99 under all conditions except for uniform parameter distributions with $N = 200$. Item discriminations correlated .927 or above, and as high as .977, except for skewed true parameter distributions with 50 items. Larger sample size resulted in higher discrimination parameter correlations except for the skewed distributions.

Table 5. Correlations Between Item Parameter Estimates From Xcalibre and Bilog for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	*a*	*b*
$N = 200$, 25 Items		
Uniform	0.972	0.984
Normal	0.928	0.993
Skewed	0.968	0.997
$N = 200$, 50 Items		
Uniform	0.927	0.988
Normal	0.962	0.999
Skewed	0.890	0.999
$N = 1{,}000$, 25 Items		
Uniform	0.994	0.995
Normal	0.943	0.999
Skewed	0.927	0.999
$N = 1{,}000$, 50 Items		
Uniform	0.977	0.996
Normal	0.977	0.999
Skewed	0.884	0.999

3PLM: No Priors/Floating Priors

The means and SDs for the *a, b*, and *c* parameters are shown in Tables 6, 7, and 8, respectively. The means for the estimated *a* parameters were all fairly close to the true means, and neither program consistently performed better than the other. The SDs of the estimated *a* parameters were smaller than the true values (again, many were close to 0.20) when Xcalibre was used but larger than the true values when Bilog was used. For the *b* parameters (Table 7), the mean estimates using Xcalibre were closer to the true means than those estimated using Bilog, for all conditions. The SDs of the *b* parameters were also closer to the true SDs for 75% of the conditions when Xcalibre was used than when Bilog was used, and the SDs from Bilog were all larger than the true values. The means for the estimated *c* parameters (Table 8) were close to the true values for both programs, but Xcalibre performed slightly better than Bilog. There were four conditions with $N = 200$ in which Bilog was not able to estimate the *c* parameters and, consequently, set all of the estimates to 0.25. The SDs of the estimated *c* parameters were

Table 6. Means and Standard Deviations of Item Discriminations for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	True		Xcalibre		Bilog	
	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	0.846	0.464	0.704	0.147	0.504	0.236
Normal	0.957	0.254	0.903	0.144	0.862	0.357
Skewed	1.186	0.204	1.001	0.130	1.181	0.382
N = 200, 50 Items						
Uniform	1.009	0.419	1.042	0.222	0.786	0.449
Normal	0.902	0.236	0.919	0.137	0.850	0.392
Skewed	1.150	0.124	1.067	0.117	1.159	0.387
N = 1,000, 25 Items						
Uniform	0.846	0.464	0.663	0.196	0.792	0.447
Normal	0.957	0.254	0.831	0.175	0.945	0.287
Skewed	1.186	0.204	1.039	0.192	1.250	0.446
N = 1,000, 50 Items						
Uniform	1.009	0.419	1.015	0.324	0.971	0.503
Normal	0.902	0.236	0.893	0.193	0.926	0.332
Skewed	1.150	0.124	1.067	0.122	1.175	0.236

Table 7. Means and Standard Deviations of Item Difficulties for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	True		Xcalibre		Bilog	
	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	−0.342	1.528	−0.516	1.404	−0.667	2.019
Normal	−0.103	0.951	−0.072	1.014	−0.060	1.617
Skewed	0.539	0.557	0.514	0.655	0.449	0.610
N = 200, 50 Items						
Uniform	−0.365	1.717	−0.405	1.521	−0.259	2.213
Normal	0.192	1.054	0.130	1.031	0.278	1.410
Skewed	0.574	0.616	0.522	0.665	0.451	0.645
N = 1,000, 25 Items						
Uniform	−0.342	1.528	−0.377	1.725	−0.568	1.728
Normal	−0.103	0.951	−0.144	0.996	−0.157	0.988
Skewed	0.539	0.557	0.517	0.652	0.421	0.666
N = 1,000, 50 Items						
Uniform	−0.365	1.717	−0.431	1.708	−0.568	1.846
Normal	0.192	1.054	0.144	1.099	0.084	1.125
Skewed	0.574	0.616	0.566	0.646	0.530	0.657

uniformly smaller than the true SDs when Xcalibre was used for estimation and larger than the true SDs when Bilog was used.

Table 8. Means and Standard Deviations of the Guessing Parameters for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	True		Xcalibre		Bilog	
	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	0.254	0.029	0.250	0.006	0.250	0.000
Normal	0.250	0.020	0.235	0.007	0.250	0.000
Skewed	0.266	0.015	0.238	0.007	0.220	0.111
N = 200, 50 Items						
Uniform	0.247	0.029	0.235	0.009	0.250	0.000
Normal	0.250	0.020	0.242	0.008	0.250	0.000
Skewed	0.256	0.012	0.235	0.007	0.208	0.114
N = 1,000, 25 Items						
Uniform	0.254	0.029	0.258	0.016	0.183	0.182
Normal	0.250	0.020	0.239	0.010	0.225	0.119
Skewed	0.266	0.015	0.250	0.014	0.228	0.082
N = 1,000, 50 Items						
Uniform	0.247	0.029	0.248	0.016	0.256	0.171
Normal	0.250	0.020	0.245	0.012	0.218	0.109
Skewed	0.256	0.012	0.258	0.013	0.252	0.080

Table 9 shows the RMSE values of the item parameter estimates. The RMSE values for the *a* parameters were generally lower for Xcalibre than for Bilog. For the *b* and *c* parameters, RMSEs were uniformly lower for Xcalibre, with substantial differences primarily for *N* = 200 conditions. Overall, the RMSE values for all three item parameters decreased when the larger sample size was used. RMSE generally decreased for the longer tests, as well.

The bias values of the item parameter estimates are shown in Table 10. For the *a* parameter estimates, neither program consistently had lower bias than the other. For the *b* and *c* parameters, however, Xcalibre had smaller values of bias than Bilog under all conditions. Most of the estimates for all three item parameters were negatively biased. Also, the bias followed the same general trend of smaller values for long tests and large sample sizes.

Table 9. RMSE of Item Parameter Estimates for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	Xcalibre			Bilog		
	a	*b*	*c*	*a*	*b*	*c*
N = 200, 25 Items						
Uniform	0.421	0.460	0.028	0.514	1.019	--
Normal	0.191	0.192	0.025	0.380	0.796	--
Skewed	0.249	0.174	0.031	0.346	0.264	0.115
N = 200, 50 Items						
Uniform	0.272	0.354	0.031	0.490	0.714	--
Normal	0.178	0.281	0.022	0.363	0.583	--
Skewed	0.166	0.173	0.028	0.397	0.288	0.128
N = 1,000, 25 Items						
Uniform	0.340	0.315	0.031	0.284	0.506	0.181
Normal	0.179	0.109	0.023	0.174	0.236	0.122
Skewed	0.180	0.119	0.019	0.339	0.228	0.087
N = 1,000, 50 Items						
Uniform	0.161	0.242	0.030	0.459	0.553	0.171
Normal	0.114	0.131	0.023	0.229	0.311	0.117
Skewed	0.140	0.089	0.012	0.227	0.176	0.074

Table 10. Bias of Item Parameter Estimates for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	Xcalibre			Bilog		
	a	*b*	*c*	*a*	*b*	*c*
N = 200, 25 Items						
Uniform	−0.142	−0.175	−0.004	−0.342	−0.325	--
Normal	−0.044	0.031	−0.015	−0.084	0.043	--
Skewed	−0.185	−0.025	−0.028	−0.005	−0.089	−0.045
N = 200, 50 Items						
Uniform	0.033	−0.040	−0.012	−0.223	0.106	--
Normal	0.016	−0.062	−0.008	−0.053	0.086	--
Skewed	−0.083	−0.052	−0.025	0.009	−0.123	−0.051
N = 1,000, 25 Items						
Uniform	−0.183	−0.036	0.004	−0.054	−0.226	−0.071
Normal	−0.116	−0.041	−0.011	−0.002	−0.054	−0.025
Skewed	−0.147	−0.022	−0.015	0.064	−0.118	−0.037
N = 1,000, 50 Items						
Uniform	0.007	−0.066	0.001	−0.038	−0.203	0.009
Normal	−0.009	−0.047	−0.005	0.024	−0.107	−0.032
Skewed	−0.083	−0.008	−0.001	0.025	−0.044	−0.007

The correlations between item parameter estimates from the two programs are shown in Table 11. The correlations between *b* parameter estimates were again very high (all above 0.92), but the correlations for the *a* parameter estimates were quite a bit lower than those obtained using the 2PLM, with correlations for $N = 200$ generally lower than for $N = 1{,}000$. The correlations for the *c* parameter estimates were lower still, suggesting that the *c* parameter estimates differed considerably between the two programs, although the low correlations could be partially due to their low variability.

Table 11. Correlations Between Item Parameter Estimates From Xcalibre and Bilog for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	*a*	*b*	*c*
$N = 200$, 25 Items			
Uniform	0.895	0.921	--
Normal	0.798	0.957	--
Skewed	0.781	0.938	0.363
$N = 200$, 50 Items			
Uniform	0.617	0.975	--
Normal	0.721	0.971	--
Skewed	0.777	0.938	0.558
$N = 1{,}000$, 25 Items			
Uniform	0.728	0.949	0.328
Normal	0.847	0.966	0.493
Skewed	0.862	0.977	0.592
$N = 1{,}000$, 50 Items			
Uniform	0.642	0.974	0.321
Normal	0.912	0.978	0.437
Skewed	0.764	0.977	0.763

3PLM: Fixed Priors

The means and SDs of the *a*, *b*, and *c* parameters are shown in Tables 12, 13, and 14, respectively. The mean *a* parameters estimated using Bilog were closer to the true means than those estimated using Xcalibre. The SDs of the *a* parameter estimates were closer to the true values using Bilog, as well. The means and SDs of the *b* parameter estimates (Table 13) were very close to the true values for both programs; neither outperformed the other. The mean *c* parameter estimates were also very close to the true values for both programs. The SDs of the estimated *c* parameters were all fairly close to the true values, with the SDs from Xcalibre always slightly smaller than the true values and the SDs from Bilog always slightly larger than the true values.

Table 12. Means and Standard Deviations of Item Discriminations for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	True		Xcalibre		Bilog	
	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	0.846	0.464	0.740	0.156	0.827	0.301
Normal	0.957	0.254	0.855	0.133	0.983	0.322
Skewed	1.186	0.204	0.901	0.116	1.102	0.238
N = 200, 50 Items						
Uniform	1.009	0.419	0.908	0.187	0.994	0.356
Normal	0.902	0.236	0.854	0.128	0.976	0.269
Skewed	1.150	0.124	0.910	0.102	1.108	0.213
N = 1,000, 25 Items						
Uniform	0.846	0.464	0.692	0.206	0.830	0.334
Normal	0.957	0.254	0.831	0.180	0.940	0.233
Skewed	1.186	0.204	0.982	0.174	1.195	0.312
N = 1,000, 50 Items						
Uniform	1.009	0.419	0.945	0.297	1.042	0.404
Normal	0.902	0.236	0.875	0.192	0.940	0.266
Skewed	1.150	0.124	0.964	0.118	1.131	0.175

Table 13. Means and Standard Deviations of Item Difficulties for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	True		Xcalibre		Bilog	
	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	−0.342	1.528	−0.495	1.378	−0.475	1.358
Normal	−0.103	0.951	−0.057	1.047	−0.101	1.120
Skewed	0.539	0.557	0.507	0.678	0.455	0.601
N = 200, 50 Items						
Uniform	−0.365	1.717	−0.417	1.610	−0.481	1.753
Normal	0.192	1.054	0.124	1.064	0.122	1.115
Skewed	0.574	0.616	0.528	0.700	0.502	0.650
N = 1,000, 25 Items						
Uniform	−0.342	1.528	−0.385	1.683	−0.318	1.560
Normal	−0.103	0.951	−0.130	1.002	−0.111	0.934
Skewed	0.539	0.557	0.499	0.661	0.454	0.604
N = 1,000, 50 Items						
Uniform	−0.365	1.717	−0.455	1.758	−0.460	1.768
Normal	0.192	1.054	0.141	1.109	0.144	1.082
Skewed	0.574	0.616	0.539	0.658	0.538	0.621

Table 14. Means and Standard Deviations of the Guessing Parameters for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	True		Xcalibre		Bilog	
	Mean	SD	Mean	SD	Mean	SD
N = 200, 25 Items						
Uniform	0.254	0.029	0.251	0.006	0.267	0.033
Normal	0.250	0.020	0.243	0.008	0.241	0.325
Skewed	0.266	0.015	0.241	0.007	0.245	0.025
N = 200, 50 Items						
Uniform	0.247	0.029	0.244	0.010	0.247	0.036
Normal	0.250	0.020	0.245	0.008	0.252	0.036
Skewed	0.256	0.012	0.240	0.007	0.235	0.034
N = 1,000, 25 Items						
Uniform	0.254	0.029	0.254	0.017	0.259	0.052
Normal	0.250	0.020	0.245	0.011	0.251	0.044
Skewed	0.266	0.015	0.243	0.012	0.245	0.038
N = 1,000, 50 Items						
Uniform	0.247	0.029	0.245	0.015	0.248	0.045
Normal	0.250	0.020	0.245	0.012	0.247	0.037
Skewed	0.256	0.012	0.246	0.011	0.258	0.043

The RMSE values of the item parameter estimates are shown in Table 15. For the *a* parameters, both programs resulted in similar values of RMSE, with no clear trend across conditions. Xcalibre generally had slightly smaller RMSE values for the *b* parameter and had uniformly smaller RMSEs for the *c* parameters. In general, a large sample size and long test resulted in smaller RMSE values than a small sample size and short test. The values of RMSE for the estimates from Xcalibre were similar to those in the floating priors condition, but the RMSEs for Bilog estimates decreased substantially from those in the no prior condition.

Table 16 shows the bias values of the item parameter estimates. For the *a* parameters, Bilog generally had smaller values of bias than Xcalibre, and Xcalibre resulted in estimates that were uniformly negatively biased. For the *b* and *c* parameters, neither program resulted in consistently lower bias than the other. In general, bias followed the same trend as RMSE; a long test and large sample size resulted in smaller values of bias. The Bilog estimates also were less biased with fixed priors than those in the no prior condition (particularly for the *b* and *c* parameters).

Table 15. RMSE of Item Parameter Estimates for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	Xcalibre			Bilog		
	a	*b*	*c*	*a*	*b*	*c*
N = 200, 25 Items						
Uniform	0.408	0.458	0.027	0.382	0.533	0.041
Normal	0.211	0.212	0.021	0.269	0.295	0.036
Skewed	0.331	0.191	0.029	0.237	0.176	0.040
N = 200, 50 Items						
Uniform	0.305	0.325	0.029	0.333	0.381	0.041
Normal	0.184	0.286	0.021	0.255	0.380	0.041
Skewed	0.276	0.189	0.023	0.234	0.188	0.044
N = 1,000, 25 Items						
Uniform	0.318	0.290	0.031	0.221	0.222	0.052
Normal	0.180	0.108	0.021	0.137	0.120	0.046
Skewed	0.226	0.131	0.025	0.197	0.138	0.039
N = 1,000, 50 Items						
Uniform	0.181	0.252	0.029	0.194	0.312	0.052
Normal	0.116	0.136	0.022	0.173	0.166	0.040
Skewed	0.218	0.099	0.017	0.167	0.121	0.037

Table 16. Bias of Item Parameter Estimates for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	Xcalibre			Bilog		
	a	*b*	*c*	*a*	*b*	*c*
N = 200, 25 Items						
Uniform	−0.106	−0.154	−0.003	−0.020	−0.133	0.012
Normal	−0.092	0.046	−0.007	0.036	0.002	−0.010
Skewed	−0.285	−0.031	−0.025	−0.084	−0.084	−0.031
N = 200, 50 Items						
Uniform	−0.101	−0.051	−0.003	−0.015	−0.116	0.000
Normal	−0.049	−0.068	−0.005	0.073	−0.069	0.002
Skewed	−0.240	−0.047	−0.019	−0.042	−0.072	−0.024
N = 1,000, 25 Items						
Uniform	−0.153	−0.044	0.000	−0.015	0.024	0.004
Normal	−0.116	−0.027	−0.006	−0.006	−0.008	0.001
Skewed	−0.203	−0.040	−0.023	0.009	−0.085	−0.021
N = 1,000, 50 Items						
Uniform	−0.064	−0.090	−0.002	0.034	−0.095	0.001
Normal	−0.028	−0.050	−0.005	0.038	−0.048	−0.003
Skewed	−0.186	−0.035	−0.013	−0.019	−0.036	−0.002

Table 17 shows the correlations between the item parameter estimates from the two programs. As in the other conditions, the highest correlations were for the *b* parameter estimates – all were above 0.988. The correlations for the *a* and *c* parameters were high as well; they increased considerably from the condition with no priors/floating priors.

Table 17. Correlations Between Item Parameter Estimates for Uniform, Normal, and Skewed True Parameter Distributions

Condition and Distribution	*a*	*b*	*c*
N = 200, 25 Items			
Uniform	0.870	0.991	0.918
Normal	0.934	0.989	0.818
Skewed	0.975	0.997	0.871
N = 200, 50 Items			
Uniform	0.878	0.993	0.884
Normal	0.856	0.991	0.877
Skewed	0.896	0.995	0.888
N = 1,000, 25 Items			
Uniform	0.920	0.997	0.939
Normal	0.945	0.994	0.887
Skewed	0.892	0.995	0.848
N = 1,000, 50 Items			
Uniform	0.967	0.995	0.843
Normal	0.956	0.992	0.902
Skewed	0.824	0.992	0.902

Discussion and Conclusions

In all of the conditions, the mean item parameter estimates were close to the true values. In the 2PLM and 3PLM (with no or floating priors) conditions, the estimated means from Xcalibre were generally closer to their true values than those of Bilog, especially for the *b* parameter. Xcalibre had lower bias and RMSE than Bilog under most conditions for the *b* parameter under the 2PLM, and for the 3PLM with no or floating priors had uniformly lower RMSE and bias for the *b* and *c* parameters, with mixed results for the *a* parameter. In the 3PLM, Xcalibre also resulted in mean *c* parameter estimates that were closer to their true values, and SDs of the *c* parameter estimates that better reflected the true values; SDs of the Bilog *c* estimates were considerably larger than the true SDs. When fixed priors were used for the 3PLM, Bilog results were generally closer to the true values for the *a* parameter (because Bilog's prior *a* mean was .3 higher than that of Xcalibre) and results were mixed for the *b* parameter. Xcalibre RMSEs for the *c* parameter were generally closer to the true values, Xcalibre RMSEs were uniformly lower for *c*, and were generally lower for the *b* parameter.

The *b* parameters were estimated relatively well regardless of the condition. The correlations between *b* parameter estimates from the two programs were very high. The *a* parameters were not estimated as well as the *b* parameters. In the 2PLM and 3PLM (with fixed priors) conditions, the *a* parameter estimates from the two programs were highly correlated, but in the 3PLM condition with no or floating priors, the correlations were reduced. That reduction was likely due to the fact that Bilog was unable to estimate the *c* parameter under several conditions for the 3PLM with no or floating priors, which would affect the other parameter estimates.

The *c* parameter estimates had the lowest correlations between Xcalibre and Bilog but most of the estimated and true values of the *c* parameter were close to 0.25 (which was 1 divided by the number of alternatives). The correlations might have been higher if there had been a wider range of *c* parameters.

For all three item parameters, estimates were generally improved when large sample sizes were used and when long tests were used. When using the 3PLM, the decision of whether to use floating or fixed priors did not make much of a difference in Xcalibre; Bilog was more affected by the use of floating priors. The estimates from Bilog were improved considerably when using fixed priors in the 3PLM.

Possibly the major limitation of this study was that different prior distributions were used in the two programs. If there had been a way to make the priors identical, the comparison of the two programs would have been clearer. As it stands, it is difficult to attribute differences in estimates to the different programs or to different prior distributions. Using no priors and floating priors in the same condition also was not ideal. However, comparing the default options of different programs to one another does have some value, since many users are likely to select the default options when estimating their own item parameters.

This study used a beta (pre-release) version of Xcalibre. Before the final version (4.1) of Xcalibre was released, some adjustments were made to the EM algorithm to improve the estimates. Parameter recovery studies with the released version of Xcalibre 4.1 were reported by Guyer and Thompson (2011b). One set of their conditions, using the 3PLM with floating priors, was similar to the normal distribution conditions reported in the present study (N = 1,000, 50 items). Results for bias in the *a* parameters showed mean bias of −0.009 in the present study versus 0.076 in the previous study, −0.047 versus 0.013 for the *b* parameter, and −0.005 versus 0.004 for the *c* parameter. RMSEs were 0.178 versus 0.117 for *a*, 0.109 versus 0.120 for b, and 0.023 versus 0.019 for *c*. Thus, the results of this study are applicable to the released version of Xcalibre and in some instances slightly underestimate the recovery of the true parameters.

References

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation if item parameters: Application of an EM algorithm. *Psychometrika, 40*, 443-459.

Guyer, R., & Thompson, N.A. (2011a). *User's Manual for Xcalibre item response theory calibration software, version 4.1.3* St. Paul MN: Assessment Systems Corporation.

Guyer, R., & Thompson, N.A. (2011b). *Item response theory parameter recovery using Xcalibre 4.1.* St. Paul MN: Assessment Systems Corporation.

R Development Core Team. (2010). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Bilog-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Chicago: Scientific Software International.