Item Response Theory Parameter Recovery Using Xcalibre™ 4.1

Rick Guyer and Nathan Thompson

Technical Report

August, 2011

Assessment Systems Corporation
www.assess.com

Saint Paul, MN

Abstract

Item response theory (IRT) presents a powerful psychometric paradigm for developing, delivering, analyzing, and scoring assessments. To utilize IRT, assessment data must be calibrated with sophisticated software designed for that purpose. Use of IRT assumes that the software accurately estimates parameters for various IRT models; the fact that a program can produce IRT parameters does not mean they are accurate. Moreover, IRT assumes that the test data is of adequate volume, in both number of items and number of examinees. The purpose of this paper was to evaluate these assumptions with the software *Xcalibre 4.1*, utilizing a 5 (model) × 3 (test length) × 4 (sample size) monte-carlo parameter recovery study. Results indicate that *Xcalibre*'s sophisticated IRT estimation algorithm accurately recovered IRT parameters, in addition to its superiority over other IRT software in terms of quality output and user friendliness.

Item Response Theory Parameter Recovery Using Xcalibre 4.1

Item response theory (IRT) presents a powerful psychometric paradigm for developing, delivering, analyzing, and scoring assessments. To utilize IRT, a statistical analysis of test data, typically called a *calibration*, is performed with sophisticated software. An obvious assumption of IRT is that the test data is of adequate volume and quality to fit IRT models. An additional, more implicit assumption is that the software is sophisticated enough to accurately fit models when the data is adequate. The purpose of this paper was to evaluate these two assumptions.

Studies that evaluate questions such as these are called *parameter recovery studies*. The rationale is that we can generate data sets with known IRT parameters (IRT is a strong enough theory that it can do so quite easily, using monte-carlo simulation methods), calibrate the data under different conditions (e.g., sample size, IRT model, software program), and then compare the IRT parameters from the calibration to the known parameters that were originally generated. If calibrated parameters are quite similar to the original parameters, that is, if they recover the original parameters, the calibration can be interpreted as accurate. It is obvious that a report with such a comparison is necessary documentation for any IRT calibration software; otherwise, the quality of the software is unknown, and it could be providing very inaccurate analyses.

There are two types of parameter recovery studies. First, initial studies investigate a newly developed IRT model and provide support for the validity of its item parameter calibration algorithm (e.g., Muraki, 1992). Later studies typically provide a comparison of different models, different sample sizes, or different software (Reise & Yu, 1990; French & Dodd, 1999; DeMars, 2004; Wang & Chen, 2005; Jurich & Goodman, 2009). The current study falls into the latter category, as the IRT models are established and well known, and the purpose was to evaluate sample size needs for calibration with new IRT calibration software, *Xcalibre 4.1* (Guyer & Thompson, 2011).

Yoes (1995) evaluated the parameter recovery of an earlier version of *Xcalibre* while also comparing it to existing programs at the time (BILOG, LOGIST, and ASCAL). *Xcalibre* was found to be the most accurate program, especially with small sample sizes or test lengths. The new version of *Xcalibre* (4.1 at this time) has received enhancements to the item parameter estimation algorithm, as well as years' worth of improvements to a user-friendly interface and output. Therefore, this study was designed to evaluate the new program under similar conditions but with expanded model comparisons in place of software comparisons.

Most notably, this study compared the three major dichotomous models, the 3-parameter logistic, 2-parameter logistic, and the Rasch (1-parameter) models, as well as Samejima's (1972) graded response model and Muraki's (1992) generalized partial credit model. The previous study evaluated only the 3-parameter model, focusing on a comparison of different software.

Like Yoes (1995), the effect of the number of items on the test (50, 100, or 200) on item parameter recovery was examined. In addition, the effect of sample size (300, 500, 1,000, and 2,000) on item parameter recovery was investigated. French and Dodd (1999) noted that smaller sample sizes can be used for Rasch models, but this study focused primarily on non-Rasch models because they do not make the unrealistic assumption of equivalent item discrimination.

Therefore, this parameter recovery study had a 5 (model) × 3 (test length) × 4 (sample size) design, providing a complex evaluation of important variables in IRT calibration. Results indicate that *Xcalibre*'s sophisticated IRT estimation algorithm accurately recovers IRT parameters. In addition to its superior accuracy, *Xcalibre* has been designed to have superior output quality and user-friendliness over other programs. Its interface is purely point-and-click, with no writing of complex idiosyncratic command code required. While output from other programs might be ASCII text or simple lists of numbers, *Xcalibre* automatically builds a comprehensive report document with dozens of embedded tables and graphics. *Xcalibre* therefore represents the most sophisticated IRT program available to researchers and practitioners.

Method

The goal of this study was to examine item parameter recovery under a range of important conditions, outlined by the 5 (IRT model) × 3 (test length) × 4 (sample size) design. The test length, or number of test items, equaled 50, 100, or 200; longer tests typically provide more reliable measurement and therefore more accurate calibration. The sample sizes used were 300, 500, 1,000, and 2,000 examinees; again, higher numbers are expected to provide more accurate calibration.

Five IRT models were evaluated. The 3-parameter logistic model (3PL), the 2-parameter logistic (2PL), and the 1-parameter (1PL, Rasch) dichotomous models were examined. In addition, Samejima's graded response model (SGRM; 1972) and Muraki's generalized partial credit model (GPCM; 1992) were also examined. A model constant (D) of 1.7 was used for the non-Rasch conditions in this study.

The dichotomous item parameters were generated according to the following distributions: $a \sim N(0.80, 0.2)$, $b \sim N(0.0, 1.0)$, $c \sim N(0.25, .03)$. To ensure comparability across the three different dichotomous models, the same *b* parameters were used for all three dichotomous models. Additionally, the same *a* parameters were used for the 2PL and the 3PL models.

The polytomous *a* parameter was generated separately and was normally distributed with a mean of 1.0 and a standard deviation of 0.2. All of the polytomous items used the traditional

five-category scale. The boundary location (b_k) parameters for categories 1 to k were fixed to be –2, –1, 1, and 2 for all items.

Monte-Carlo Design

A monte-carlo design was used in this study. A total of 20 replications were performed for each of the 60 cells in the design. Simulees were generated using a standard normal distribution for each replication in each cell. The generated θ values were scaled to have a mean of 0.0 and a standard deviation of 1.0.

The generated item parameters were used for purposes of the monte-carlo simulation. A matrix of random numbers was drawn from a uniform distribution with a minimum of 0 and maximum of 1 (U[0, 1]). This random number matrix had as many rows as persons and as many columns as items. A probability matrix was generated based on θ and the IRT parameters according to the given IRT model.

The procedure of generating item responses was different for dichotomous and polytomous items. For each cell in the matrix, the random number was compared to the probability of a correct response to create a dichotomous item response. If the random number was greater than the probability, the response was a 0, while a probability greater than the random number resulted in a response of 1.

For polytomous items, the cumulative probability of responding in at least category k was computed for all categories. If the random number was less than the probability of responding in category 1, the response was coded as a 1. If the random number was greater than the cumulative probability for category k but less than the probability for category $k - 1$, then the response was coded as k. This process results in an appropriate polytomous response matrix.

Parameter Recovery

One goal of this study was to evaluate the recovery of the generated item parameters. The first index of item parameter recovery that was used was the average bias, which was defined as

$$\text{Bias} = \frac{\sum_{i=1}^{20} (\hat{\xi}_i - \xi)}{20}, \qquad (1)$$

where ξ is a given item parameter and $\hat{\xi}_i$ is its estimate.

The root mean square error (RMSE) has the advantage of being in the same metric as the item parameters. It is defined by

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{20}(\hat{\xi}_i - \xi)^2}{20}}. \quad (2)$$

The final index is the standard error. It was defined as

$$\text{SE} = \sqrt{\frac{\sum_{i=1}^{20}(\hat{\xi}_i - \bar{\hat{\xi}})^2}{20}}. \quad (3)$$

The SE equals the standard deviation of the estimated item parameters over the 20 replications performed in the study, and indexes instability in the item parameter estimates.

Correlations between the generated item parameters and the estimated item parameters were computed for each replication. The correlation was defined as

$$r = \frac{\sum_{i=1}^{n}(\hat{\xi}_i - \bar{\hat{\xi}})(\xi_i - \bar{\xi})}{(n-1)s_{\hat{\xi}}s_{\xi}}, \quad (4)$$

where n equals the number of items in the condition.

Software Specifications

The computer program *Xcalibre* Version 4.1 (Guyer & Thompson, 2011) was used for the item parameter calibration performed in this study. It uses a marginal maximum likelihood procedure with an expectation maximization (E-M) algorithm. Except for the Rasch model, all item parameter calibrations were performed using a D of 1.7. The program used an item parameter convergence criterion of 0.01 with an upper limit of 80 E-M cycles.

The following prior distributions were used (where relevant): a ~ N(1.0, 0.3); b ~ N(0, 1); c ~ N(0.25, 0.03). The prior item parameter distributions were allowed to float after the completion of each item parameter estimation step of the E-M algorithm. This means that the mean of the prior was updated to equal the mean of the estimated item parameters after the given iteration.

To ensure comparability of the item parameters across samples, *Xcalibre* provides the option to center the item parameter estimates on θ. For the 2PL, 3PL, and the polytomous models *Xcalibre* scaled the θ estimates to have a standard deviation of 1.0. This was done by multiplying the a parameters by the standard deviation of θ. For the 2PL and 3PL models, the θ

distribution is centered to have a mean of 0.0. Version 4.1 of *Xcalibre* uses the user specified θ estimation method to perform the θ estimate centering. Thus, for this simulation study maximum likelihood θ estimates were used for the item parameter centering. Simulees that did not have a reasonable θ estimate were excluded from the θ distribution for purposes of item parameter centering. For the Rasch model the *b* parameter estimates were centered to have a mean of 0.0 and standard deviation of 1.0.

Results

There were as many bias, SE, and RMSE statistics as there were items for each cell in the study. As the generating values for the *a*, *b*, and *c* parameters differed across items, the median item parameter recovery was computed and is reported in the tables. The generating value did not vary for the polytomous boundary locations so the mean recovery was computed across items.

Dichotomous Models

The bias, standard error, and RMSE statistics are presented in Table 1 for the Rasch model. It was evident that the median bias across all items was essentially zero for each of the conditions examined. The median SE and RMSE decreased as sample size increased. The recovery of the *b* parameter was not sensitive to the number of items in the test.

The results for the 2-parameter model are presented in Table 2. There was evidence that the *a* parameter estimates had a minimal amount of positive bias for shorter tests; it decreased as the number of items in the test was increased from 50 (bias of about 0.050) to 200 (bias of about 0.015). The median bias for the *b* parameter was essentially zero for all conditions in this study. The median SE decreased for the *a* and *b* parameters as the sample size increased. The RMSEs decreased as both the sample size and the number of items increased.

The *a* parameters were also positively biased for the 3-parameter model as shown by Table 3. As with the 2-parameter model, the positive bias decreased as a function of the number of items in the test. The median bias statistics for the *b* and *c* parameters were near zero. The SE and RMSE statistics decreased as sample size increased for the *a* and *b* parameters. The SEs of the *c* parameters slightly increased as sample size was increased.

Table 1. Item Parameter Recovery for the Rasch Model *b* Parameter for *Xcalibre 4.1*, for *N* = 300, 500, 1,000, and 5,000

	Bias				SE				RMSE			
No. Items	300	500	1000	2000	300	500	1000	2000	300	500	1000	2000
50	.006	.001	−.001	.001	.142	.119	.092	.062	.155	.134	.106	.070
100	.000	-.012	.002	−.001	.147	.118	.091	.060	.158	.128	.099	.079
200	−.004	−.001	−.002	.001	.151	.118	.086	.060	.171	.136	.107	.085

Table 2. Item Parameter Recovery for the 2PL for *Xcalibre 4.1*, for *N* = 300, 500, 1,000, and 5,000

No. Items &	Bias				SE				RMSE			
Parameter	300	500	1000	2000	300	500	1000	2000	300	500	1000	2000
50 Items												
a	.054	.057	.049	.053	.090	.079	.059	.040	.116	.101	.081	.067
b	.000	−.001	.012	.006	.114	.093	.062	.044	.124	.099	.081	.062
100 Items												
a	.031	.026	.032	.028	.088	.075	.058	.041	.102	.086	.068	.050
b	.000	.002	.004	.004	.115	.093	.065	.047	.119	.098	.071	.053
200 Items												
a	.018	.015	.014	.015	.086	.073	.053	.040	.097	.081	.060	.045
b	.002	−.003	−.001	.001	.108	.086	.062	.044	.109	.089	.064	.046

Table 3. Item Parameter Recovery for the 3PL for *Xcalibre 4.1*, for N = 300, 500, 1,000, and 5,000

No. Items &	Bias				SE				RMSE			
Parameter	300	500	1000	2000	300	500	1000	2000	300	500	1000	2000
50 Items												
a	.089	.093	.076	.089	.110	.102	.080	.066	.150	.141	.117	.109
b	.004	.013	.013	−.007	.139	.114	.089	.067	.167	.138	.120	.109
c	.007	.005	.004	.002	.006	.007	.008	.008	.021	.020	.019	.020
100 Items												
a	.061	.057	.050	.050	.108	.089	.077	.060	.138	.121	.103	.082
b	.021	.019	.009	.010	.147	.123	.083	.065	.163	.140	.112	.086
c	.004	.004	.001	.002	.006	.007	.009	.010	.023	.023	.022	.021
200 Items												
a	.030	.032	.028	.030	.106	.094	.076	.058	.128	.113	.089	.069
b	−.007	−.009	−.003	−.002	.145	.116	.081	.063	.166	.138	.104	.085
c	−.004	−.003	−.003	−.002	.006	.007	.008	.010	.023	.022	.021	.021

Polytomous Models

Bias

The bias statistics for the polytomous models are provided in Table 4. The *a* parameter estimates remained essentially unbiased for the SGRM and GPCM for all sample sizes and number of items used. The bias of the boundary location parameters remained quite close to zero for the SGRM. The bias increased for the boundary parameters for the SGRM 200-item condition with a sample size of 300.

The bias of the boundary location parameters for the GPCM also remained near zero across sample sizes when there were 50 items in the test. For the 100-item test there was some slight bias present when the sample sizes were 300 and 500. This bias disappeared when the sample size was increased to 1,000. The bias of the boundary location parameters increased when the number of items was 200. As before, the magnitude of the bias decreased as the sample size was increased.

Table 4. Item Parameter Bias for the Polytomous Items
for *Xcalibre 4.1*, for N = 300, 500, 1,000, and 5,000

No. Items &	SGRM				GPCM			
Parameter	300	500	1000	2000	300	500	1000	2000
50 Items								
a	−.006	.001	.000	−.002	.005	.023	.009	.002
b_1	−.007	−.010	.015	.019	−.016	.030	.011	.006
b_2	.005	−.008	.012	.018	.009	.032	.013	.019
b_3	.003	−.023	−.004	.000	.004	−.019	−.013	.002
b_4	.015	−.023	−.007	−.002	.011	−.017	−.015	.000
100 Items								
a	.008	−.008	−.008	−.002	−.001	−.003	−.007	−.011
b_1	.007	−.028	−.013	−.001	.020	.020	−.011	−.014
b_2	.010	−.013	−.001	.002	.029	.026	.001	−.003
b_3	−.003	−.003	−.005	−.007	.040	.032	−.004	.001
b_4	−.002	.012	−.013	−.007	.040	.040	.006	.010
200 Items								
a	−.012	−.005	.009	−.008	.006	−.016	−.013	−.011
b_1	−.029	−.025	.006	−.021	.073	−.025	−.009	−.011
b_2	−.002	−.012	.001	−.012	.081	.002	.008	−.001
b_3	−.035	.003	−.015	−.003	.057	.019	.015	.006
b_4	.059	.015	−.019	.006	.065	.036	.025	.011

SE

The median SEs for the *a* parameter decreased as sample size increased for the SGRM and GPCM, as shown by Table 5. The *a* parameter SEs were not sensitive to the number of items in the test. The SEs for boundaries 1 and 4 were larger than the SEs for boundaries 2 and 3. The SEs of the boundary parameters also decreased as sample size increased. No systematic results were found for number of items in the test, as the SEs fluctuated across boundary parameter and sample size.

Table 5. Item Parameter SE for the Polytomous Items for *Xcalibre 4.1*, for N = 300, 500, 1,000, and 5,000

No. Items &	SGRM				GPCM			
Parameter	300	500	1000	2000	300	500	1000	2000
50 Items								
a	.085	.069	.049	.038	.086	.073	.054	.040
b_1	.201	.160	.109	.079	.226	.168	.122	.080
b_2	.134	.100	.066	.049	.131	.104	.075	.047
b_3	.114	.094	.068	.046	.133	.095	.077	.050
b_4	.190	.150	.109	.077	.218	.161	.119	.084
100 Items								
a	.085	.065	.046	.033	.081	.070	.051	.036
b_1	.190	.154	.104	.076	.220	.175	.122	.082
b_2	.118	.102	.066	.049	.145	.115	.081	.055
b_3	.134	.095	.068	.049	.166	.110	.082	.052
b_4	.205	.150	.110	.076	.239	.178	.125	.084
200 Items								
a	.081	.067	.047	.035	.084	.066	.050	.038
b_1	.195	.150	.106	.075	.233	.190	.126	.096
b_2	.128	.094	.071	.048	.152	.126	.084	.069
b_3	.136	.106	.071	.054	.133	.118	.087	.060
b_4	.212	.165	.106	.083	.218	.176	.131	.087

RMSE

The results for the RMSE mirrored those found for the SE, as seen in Table 6. The RMSEs decreased as sample size increased. In addition, the SGRM had lower RMSEs than the GPCM – a result that was most evident when the sample size was 500 or less.

Table 6. Item Parameter RMSE for the Polytomous Items for *Xcalibre 4.1*, for N = 300, 500, 1,000, and 5,000

No. Items &	SGRM				GPCM			
Parameter	300	500	1000	2000	300	500	1000	2000
50 Items								
a	.089	.074	.050	.039	.094	.080	.057	.042
b_1	.201	.160	.110	.082	.227	.171	.123	.080
b_2	.134	.101	.067	.052	.131	.109	.076	.051
b_3	.114	.096	.068	.046	.133	.097	.078	.050
b_4	.191	.152	.109	.077	.218	.162	.120	.084
100 Items								
a	.088	.068	.047	.035	.086	.073	.053	.038
b_1	.190	.157	.105	.076	.221	.176	.123	.083
b_2	.118	.103	.066	.049	.148	.118	.081	.055
b_3	.134	.095	.069	.050	.170	.115	.082	.052
b_4	.205	.151	.111	.077	.243	.182	.125	.084
200 Items								
a	.086	.070	.050	.036	.090	.072	.054	.040
b_1	.198	.153	.106	.077	.244	.192	.126	.096
b_2	.128	.095	.071	.049	.172	.126	.084	.069
b_3	.140	.106	.072	.054	.144	.120	.088	.060
b_4	.220	.166	.108	.083	.228	.180	.134	.088

Correlations

The median correlations between the item parameters are presented in Table 7. The values presented in Table 7 are the medians computed across the 20 replications. The median correlations for the *b* parameter ranged from .973 to .998 for the dichotomous IRT models. The correlations for the *a* and *c* parameters increased in strength as sample size increased. The *a* and *b* parameter correlations were lower for the 3-parameter model than they were for the other models.

The *a* parameter correlations for the SGRM and GPCM were similar in magnitude to those from the dichotomous correlations. As with the dichotomous models, the correlations increased as sample size increased. For the 300 and 500 sample size conditions, the correlations were higher for the SGRM than they were for the GPCM.

Table 7. Median Correlations Between Estimated and Generated Item Parameters for *Xcalibre 4.1*, for N = 300, 500, 1,000, and 5,000

Model and	50 Items				100 Items				200 Items			
Parameter	300	500	1000	2000	300	500	1000	2000	300	500	1000	2000
1PL												
b	.988	.992	.995	.997	.985	.991	.994	.997	.985	.990	.995	.997
2PL												
a	.890	.919	.959	.979	.876	.913	.957	.976	.860	.906	.951	.975
b	.987	.991	.995	.998	.986	.991	.995	.998	.988	.992	.996	.998
3PL												
a	.756	.845	.903	.938	.791	.838	.897	.942	.754	.823	.892	.934
b	.973	.982	.989	.993	.976	.982	.989	.994	.976	.983	.989	.993
c	.240	.363	.420	.514	.272	.378	.394	.481	.263	.333	.394	.505
SGRM												
a	.876	.914	.958	.976	.889	.928	.964	.983	.887	.930	.965	.982
GPCM												
a	.840	.898	.954	.972	.873	.914	.957	.977	.872	.913	.956	.978

Discussion

Bias

It was found that the *a* parameter for the dichotomous models showed some slight positive bias. This bias decreased as the number of items increased. The positive bias resulted from low *a* parameters being overestimated. The *b* and *c* parameters showed no bias at the test level.

The polytomous boundary parameters did show some bias for the 300 sample size condition. However, the bias dissipated as sample size was increased. It should be noted that the bias was larger for the GPCM than the SGRM model. Recovery of the GPCM boundaries improved when there were at least 500 examinees.

Overall, *Xcalibre* provided unbiased item parameter estimates at the test level. This result is supported by the essentially zero median bias values observed for the *a*, *b*, and *c* parameter estimates. In addition, the boundary location parameters for the polytomous models were estimated with minimal bias.

SE and RMSE

The results shown in Tables 2 and 3 revealed that the SEs of the 2-parameter *a* and *b* parameter estimates were lower than they were for the 3-parameter model. This result was likely due to the addition of a guessing parameter. Any estimation error in the guessing parameter would also increase the estimation error in the other item parameters, as well. For this reason, the sample size requirement for the 3-parameter model is larger than it is for the 2-parameter model.

Overall, the *a* and *b* parameter SEs were low and decreased rapidly as sample size increased. This indicated that *Xcalibre* can provide stable item parameter estimates with low empirical SEs. The guessing parameter showed low median empirical SEs and RMSEs. This suggested that *Xcalibre* was able to estimate the *c* parameter quite well, even for a sample size of 300.

For the polytomous models it was found that the SEs and RMSEs for the SGRM were generally lower than they were for the GPCM. This result was strongest for the 300 and 500 examinee conditions. This provided evidence that the GPCM required larger sample sizes for stable estimates than the SGRM.

The SEs and RMSEs for the polytomous models indicate that the *a* parameters were recovered well. In addition, the SEs decreased rapidly with increased sample size and indicated that *Xcalibre* can provide stable estimates for samples larger than 300 examinees.

Correlations

The correlations between the item parameters followed the trend observed by Yoes (1995). The correlations were strongest for the *b* parameter, second strongest for the *a* parameter, and weakest for the *c* parameter. The magnitude of the correlations can be explained by the lower variance in the *a* and *c* parameters compared to the *b* parameter. The extremely high correlations (.973 to .998) provide evidence for the calibration accuracy of *Xcalibre 4.1*.

Conclusions

Dichotomous item parameters were recovered with little bias and low SEs, provided samples of at least 300 were used. The polytomous parameters were recovered with little bias (with samples of at least 500) and SEs that quickly decreased as sample size increased. The results of this study indicated that *Xcalibre* provides stable and unbiased item parameter estimates. However, note that the accuracy of parameter estimates increased substantially with sample sizes of 1,000 or 2,000, even for the Rasch model.

References

Demars, C. (2004). *A comparison of the recovery of parameters using the nominal response and generalized partial credit models*. Poster presented at the annual meeting of American Educational Research Association.

French, G.A. and Dodd, B.G. (1999). Parameter recovery for the rating scale model using PARSCALE. *Journal of Outcome Measurement, 3*, 176-199.

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

Jurich, D., & Goodman, J. (2009). *A comparison of IRT parameter recovery in mixed format examinations using PARSCALE and ICL*. Poster presented at the Annual meeting of Northeastern Educational Research Association.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

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

Samejima, F. (1972). *A General Model for Free-Response Data.* (Psychometric Monograph No. 18). Richmond, VA: Psychometric Society.

Wang, W-C. and Chen, C-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*, 376-404.

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model.* Saint Paul, MN: Assessment Systems Corporation.