ACT Research Report Series 2009-2

Linking Item Parameters to a Base Scale

Taehoon Kang

Nancy Petersen

ACT

August 2009

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

Linking Item Parameters to a Base Scale

Taehoon Kang
Nancy S. Petersen

Abstract

This paper compares three methods of item calibration—concurrent calibration, separate calibration with linking, and fixed item parameter calibration—that are frequently used for linking item parameters to a base scale. Concurrent and separate calibrations were implemented using BILOG-MG. The Stocking and Lord (1983) characteristic curve method of parameter linking was used in conjunction with separate calibration. The fixed item parameter calibration (FIPC) method was implemented using both BILOG-MG and PARSCALE because the method is carried out differently by the two programs. Both programs use multiple EM cycles but BILOG-MG does not update the prior ability distribution during FIPC calibration whereas PARSCALE updates the prior ability distribution multiple times. The methods were compared using simulations based on actual testing program data and results were evaluated in terms of recovery of the underlying ability distributions, the item characteristic curves, and the test characteristic curves. Factors manipulated in the simulations were sample size, ability distributions, and numbers of common (or fixed) items. The results for concurrent calibration and separate calibration with linking were comparable and both methods showed good recovery results for all conditions. Between the two fixed item parameter calibration procedures, only the appropriate use of PARSCALE consistently provided item parameter linking results similar to those of the other two methods.

Linking Item Parameters to a Base Scale

In practice, psychometricians often use item response theory (IRT) to equate new test forms or to link parameters for pretest items to a base scale. In the context of equating, the new form is often comprised of both new and old operational items where the old operational items, called common items, were previously administered in another form of the test, referred to as the old form, and test-takers' scores are based on their responses to all operational items. In the context of pretesting, some newly developed items, known as pretest items, may also be administered to test-takers along with the operational items in the test form and test-takers' scores are based on their responses to the operational items only.

In both scenarios, we need to calibrate (that is, estimate the parameters of) the new items along with the "old" (previously calibrated) operational items and then place the new item parameters onto the already established (base) scale in order either to score the test or to evaluate item quality. Several different calibration methods have been applied in practice in both the equating and pretesting scenarios, namely, concurrent calibration, separate calibration with linking, and fixed item parameter calibration (FIPC).

Many testing programs require contractors to use FIPC for the purpose of linking item parameters to a base scale. There is significant research indicating that FIPC tends to yield biased estimates (see Baldwin, Baldwin, & Nering, 2007; Keller, Keller, & Baldwin, 2007; Paek & Young, 2005; Skorupski, Jodoin, Keller, & Swaminathan, 2003). Kim (2006), however, has identified a procedure for implementing FIPC that may yield satisfactory results. This paper compares, under the equating scenario, IRT fixed item parameter calibration as traditionally implemented and as recommended in Kim with concurrent calibration and separate calibration with linking. Because IRT item parameter scaling and linking is required for many operational

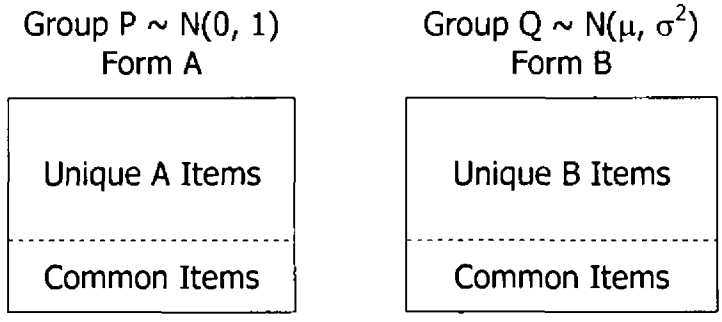
applications such as test score equating, pretest item calibration, differential item functioning, computerized adaptive testing, and so on, it is important to examine and confirm which linking procedure is most psychometrically accurate and robust.

Methods

Data and Study Design

In this study, the various calibration procedures are compared using simulated data. However, the generating item parameters were selected from calibrations of actual testing program data sets (two 60-item mathematics achievement test forms with sample sizes ranging from 4,294 to 4,557). It is assumed that we have data for two forms of a 50-item multiple-choice test and that the two forms have a set of items in common (number of common, or fixed, items equals 10, 20, or 40). It is further assumed that the data were collected via a non-equivalent groups anchor-test (NEAT) data collection design (see Figure 1). Item parameters for the old form (form A) already exist from the calibration of data from a previous administration. The task of interest here is to equate scores on the new form (form B) to those on the old form, and to do that item parameters for the new form need to be estimated (calibrated) and placed onto the scale of the old form items.

FIGURE 1. NEAT Data Collection Design for Common Item Equating



In the NEAT design, one test form is administered to one group of test-takers and another test form is administered to another group of test-takers. The two groups are naturally occurring and therefore likely to differ in ability. For example, one might be a group taking the test in the fall and the other a group taking the test in the spring. A common test or anchor test, in this case common items, is administered to both groups in order to estimate the performance of the combined group on both forms, thus simulating, by statistical methods, the situation in which both groups take both forms.

The three-parameter-logistic (3PL) IRT model is used and the ability distribution of the base or reference group (group P) is assumed to be N(0,1). Other factors included in the simulation design were (1) two different sample sizes for both the base and target groups (500 and 2000) and (2) three different ability distributions for the target group [i.e., group Q~ N(0,1), $N(.25,1.1^2)$, or $N(.5,1.2^2)$]. These factors are critical ones that affect calibration in practice.

The sample size of 2000 was chosen to represent the usual practice of calibrating operational items with relatively large samples in order to produce stable item parameter estimates. The sample size of 500 was chosen to represent the minimum sample size in practice that is likely to yield acceptable calibration results with a 3PL model. To generate old form group (group P) examinees (simulees) a N(0,1) distribution was used. The new form group (group Q) distributions were selected to represent those situations where the old and new form groups are very similar in ability [N(0,1)], differ somewhat in ability [$N(.25,1.1^2)$], and differ significantly in ability [$N(.5,1.2^2)$]. In practice, we seldom see group differences as extreme as those represented by the $N(.5,1.2^2)$ distribution.

There are a total of 18 conditions simulated in this study (2 sample sizes *x* 3 numbers of fixed items *x* 3 target group ability distributions). One hundred replications were generated for each condition for both base and target groups.

IRT Calibration and Linking Methods

When concurrent calibration is used with the NEAT design for equating, item parameters for the operational items in both the new and the old forms are estimated simultaneously in a single calibration run. Because the new and old forms have items in common, the resulting item parameters for all items in the concurrent calibration run are on the same scale. In every concurrent calibration run for this study, the old form group was treated as the reference group having 0 and 1 as the mean and standard deviation (SD), respectively, of their ability estimates. The mean and SD of the new form (target) group's ability distribution were newly estimated through the item and ability parameter calibration process for each dataset. By treating the old form group as the reference group, instead of the pooled (old plus new) group, the parameters for both groups are placed onto the scale for the old form group.

In separate calibration with linking, the items taken by each group are calibrated in two separate runs, one for each group (groups P and Q in Figure 1). Using the common items as linking items, a linear transformation is then estimated to place the item parameters for the new form group, group Q, onto the scale for the old form group, group P. The linear transformation can be estimated using a variety of linking methods (e.g., see Marco, 1977; Loyd & Hoover, 1980; Haebara, 1980; and Stocking & Lord, 1983). Both concurrent and separate calibrations were implemented using BILOG-MG. And, the Stocking and Lord characteristic curve method for parameter linking was used in this study.

In FIPC, items are again calibrated in two separate runs but, unlike with separate calibration, there is no linking step. Instead, in the new form group calibration run, items are calibrated with the item parameters for the common items fixed at their separately estimated (old form) values so that the parameters for the new items are placed on the same scale as that for the old operational items. Kim (2006) compared five FIPC procedures that differ from one another according to (1) how many times they update the prior ability distribution and (2) how many EM (Expectation-Maximization) cycles they use. Kim recommended updating the prior ability distribution multiple times and the use of multiple EM cycles during the calibration process [that is, the multiple weights updating and multiple EM cycles (MWU-MEM) method] because only this approach worked well regardless of the different ability distributions used for the target group in his study. Kim's recommended FIPC procedure is used in this study.

For the purpose of comparison, another FIPC approach without the prior update is also considered in this study. In this procedure, the prior ability distribution is not updated after each M step in the EM cycles, which may result in the mean and standard deviation of the newly estimated item parameters tending to shrink toward the N(0,1) scale rather than following the metric of the fixed item parameters.

The fixed parameter calibration procedure was implemented using both BILOG-MG and PARSCALE because the procedure is carried out differently by the two programs. The goal of FIPC is to put the newly estimated item parameters on the scale of the common items that were "fixed" during calibration. Both BILOG-MG and PARSCALE use multiple EM cycles. But in BILOG-MG, use of the EMPIRICAL command which enables updating of the prior ability distribution overrides use of the NOADJUST command which prevents rescaling of parameters. Given it is more important during FIPC to not rescale parameters than to update the prior ability

distribution, BILOG-MG was implemented using only the NOADJUST command which prevents any rescaling toward mean=0 and SD=1. This problem does not occur with use of PARSALE. In PARSCALE, the NOADJUST command which prevents rescaling can be used along with the POSTERIOR command which enables updating of the prior ability distribution multiple times as recommended by Kim (2006). Table 1 provides a concise description of FIPC as implemented by BILOG-MG and PARSCALE. Examples of the BILOG-MG and PARSCALE control cards used for FIPC calibration are given in the Appendix.

TABLE 1

Fixed Item Parameter Calibration (FIPC)

<table>
<tr><td>Goal of FIPC</td><td colspan="2">The newly estimated item parameters for unique items in form B should be on the scale of the fixed or common items in form A.</td></tr>
<tr><td>Two Essential Elements of the MMLE-EM Approach for FIPC</td><td>No Rescaling:
For FIPC, do not solve the indeterminacy problem in the IRT analysis by rescaling the mean and standard deviation of the posterior ability distribution to be 0 and 1, respectively.</td><td>Update the Ability Prior after each M Step through the EM Algorithm:
By updating the prior ability distribution after each M step through the EM algorithm, the prior scale will gradually become closer to that of the fixed items in form A.</td></tr>
<tr><td rowspan="2">Related Options in BILOG-MG</td><td>NOADJUST in the CALIB command prevents any rescaling.</td><td>EMPIRICAL in the CALIB command enables use of the empirical/posterior distribution after each M step as the prior for the next EM cycle.</td></tr>
<tr><td colspan="2">Unfortunately, in the CALIB command of BILOG-MG, NOADJUST and EMPRICAL cannot be used together. As Kim (2006) indicated, when EMPIRICAL is used with NOADJUST, the latter becomes invalid. If only NOADJUST is used, automatically, the prior does not change during the EM cycles.</td></tr>
<tr><td rowspan="2">Related Options in PARSCALE</td><td>FREE=(NOADJUST, NOADJUST) in the CALIB command prevents any rescaling.</td><td>POSTERIOR in the CALIB command enables use of the empirical/posterior distribution after each M step as the prior for the next EM cycle.</td></tr>
<tr><td colspan="2">In the CALIB command of PARSCALE, FREE=(NOADJUST, NOADJUST) and POSTERIOR can be used together.</td></tr>
</table>

Evaluation Criteria

The accuracy or performance of the four IRT linking procedures used in this study was evaluated, first, in terms of recovery of the underlying ability distributions (i.e., the true mean and SD of the target group's ability distribution). Also, the ability of the four procedures to recover the item parameters was evaluated using the item characteristic curve (ICC) criterion

based on Hanson and Béguin (2002) and the test characteristic curve (TCC) criterion shown in Equation 1. The latter is similar to the former except it uses TCC curves instead of ICC curves. In this paper, when the ICC criterion was calculated for each linking method and each condition, only non-common items were used. But, for the TCC criterion, all 50 items (both common and unique items) were used. The ICC criterion was used to evaluate the ability of the procedures to recover the item parameters rather than a comparison of the parameter values from the various procedures as different sets of a-, b-, and c-parameters can produce comparable ICCs.

The TCC criterion evaluates how close the estimated TCCs are to the true TCC for the new form. The TCC criterion for each condition is given by

Equation 1:

$$\frac{1}{100}\sum_{r=1}^{100}\int_{-\infty}^{\infty}[\tau(\theta)-t_r(\theta)]^2 f(\theta)d\theta =$$

$$\int_{-\infty}^{\infty}[\tau(\theta)-\bar{t}(\theta)]^2 f(\theta)d\theta+\frac{1}{100}\sum_{r=1}^{100}\int_{-\infty}^{\infty}[t_r(\theta)-\bar{t}(\theta)]^2 f(\theta)d\theta,$$

where τ is the true TCC using the generating item parameters, t_r is the TCC using the estimated item parameters from replication r, $\bar{t}(\theta)=\frac{1}{100}\sum_{r=1}^{100}t_r(\theta)$, and $f(\theta)$ is the N(0,1) density for ability θ. The left-hand term of Equation 1 is the mean squared error (MSE) of the estimated TCCs for a condition, which can be decomposed into the squared bias (SB) and variance (VAR) as shown on the right-hand side of the equation. A Monte Carlo integration algorithm was used to evaluate the integrals in Equation 1.

Results

For each group P and Q, 100 data sets were generated for each of the 18 conditions simulated in this study, and the four different linking procedures were applied to them. When

the sample size was 500, sometimes the PARSCALE calibration runs for FIPC did not converge successfully with 3,000 EM cycles. For example, in the condition with 10 fixed or common items, 500 examinees, and a true N(0,1) ability distribution for group Q, 11 of the initial 100 replications did not converge. When calibration runs were not successful, problematic replications were discarded and other data sets were generated until 100 replications per condition without any convergence error were obtained. In the case of BILOG-MG, all calibration runs converged without any run-time error.

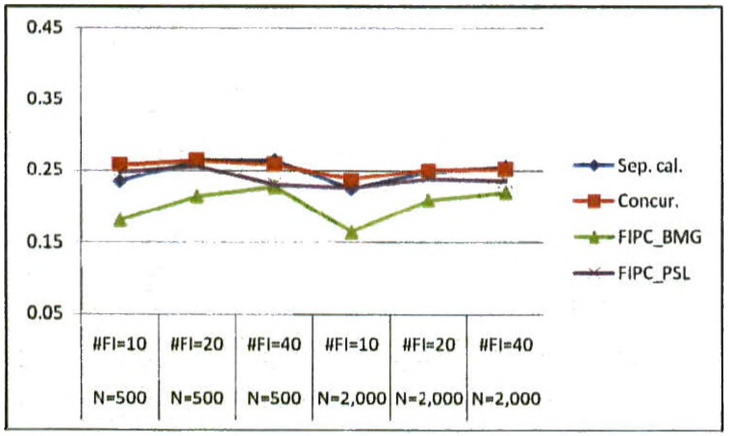
Table 2 and Figure 2 contain the results for how well the target groups' true distributions were recovered by each of the four linking procedures. When the target group (group Q in Figure 1) and the base group (group P) both had a N(0,1) ability distribution, all four linking procedures showed good and comparable recovery results and the results tended to improve slightly as the sample size increased: For sample size N=500, the average means and SDs for the target group across all the linking procedures ranged between -.02 and .03 and between .95 and 1.07, respectively. And for N=2,000, the average means and SDs ranged between -.01 and .01 and between .97 and 1.00, respectively.

TABLE 2

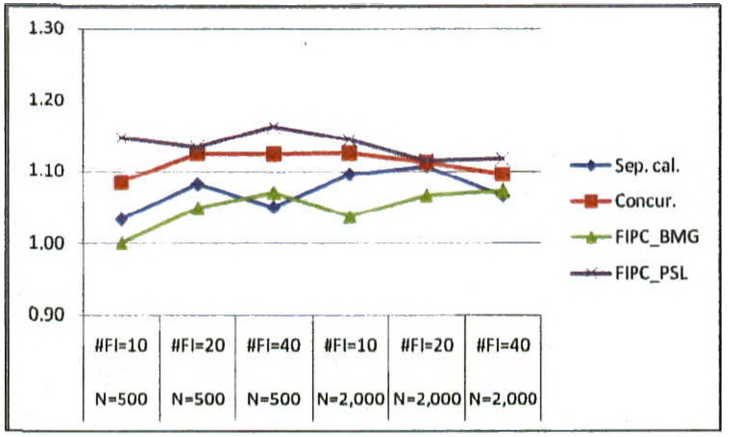
Average Means and Standard Deviations of the Estimated Ability Distributions for the Target Group

Sample Size	#Fixed or Common Items	True Target Distribution	Sep. calibration with linking		Concurrent calibration		FIPC (BILOG-MG)		FIPC (PARSCALE)	
			mean	SD	mean	SD	mean	SD	mean	SD
N=500	#FI=10	N(0.0,1.0)	0.03	0.96	0.02	0.96	0.01	0.95	0.00	1.04
		$N(.25,1.1^2)$	0.23	1.03	0.26	1.08	0.18	1.00	0.25	1.15
		$N(.50,1.2^2)$	0.51	1.11	0.52	1.16	0.37	1.04	0.54	1.27
	#FI=20	N(0.0,1.0)	0.02	0.99	0.02	1.01	0.01	0.98	-0.02	1.07
		$N(.25,1.1^2)$	0.26	1.08	0.26	1.12	0.21	1.05	0.26	1.13
		$N(.50,1.2^2)$	0.53	1.21	0.54	1.27	0.43	1.11	0.52	1.26
	#FI=40	N(0.0,1.0)	0.01	0.98	0.03	1.01	0.01	0.97	0.01	1.00
		$N(.25,1.1^2)$	0.26	1.05	0.26	1.12	0.23	1.07	0.23	1.16
		$N(.50,1.2^2)$	0.49	1.15	0.53	1.22	0.44	1.11	0.47	1.22
N=2,000	#FI=10	N(0.0,1.0)	0.00	0.99	0.00	0.97	0.00	0.97	-0.01	0.98
		$N(.25,1.1^2)$	0.22	1.10	0.24	1.13	0.16	1.04	0.23	1.14
		$N(.50,1.2^2)$	0.49	1.20	0.50	1.21	0.34	1.05	0.49	1.22
	#FI=20	N(0.0,1.0)	0.00	1.00	0.01	1.00	0.01	0.99	0.00	1.00
		$N(.25,1.1^2)$	0.25	1.11	0.25	1.11	0.21	1.07	0.24	1.11
		$N(.50,1.2^2)$	0.47	1.18	0.49	1.23	0.41	1.12	0.48	1.23
	#FI=40	N(0.0,1.0)	0.00	0.97	0.01	0.97	0.00	0.98	0.00	1.00
		$N(.25,1.1^2)$	0.25	1.07	0.25	1.10	0.22	1.07	0.24	1.12
		$N(.50,1.2^2)$	0.51	1.15	0.50	1.18	0.44	1.12	0.47	1.19

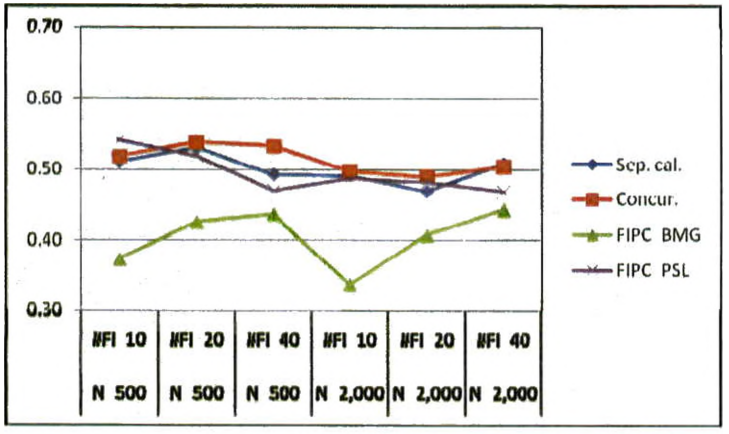
FIGURE 2. Average means and standard deviations of estimated ability distributions for target group



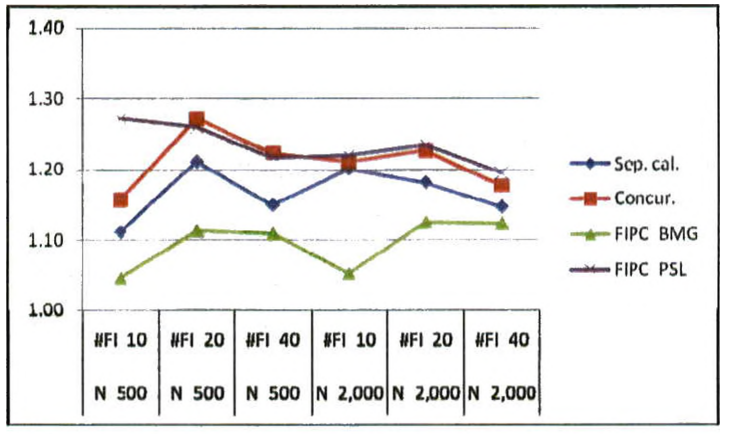
(a) Average Means when True Distribution is N(.25,1.1^2)



(b) Average SDs when True Distribution is N(.25,1.1^2)



(c) Average Means when True Distribution is N(.50,1.2^2)



(d) Average SDs when True Distribution is N(.50,1.2^2)

Figure 2 presents a summary of the results for when the true ability distribution of target group Q was either $N(.25,1.1^2)$ or $N(.5,1.2^2)$. First of all, FIPC without prior update as conducted by BILOG-MG (FIPC-BMG) consistently underestimated the true means and SDs across every condition, even though FIPC-BMG appeared to provide better recovery as the number of fixed or common items (#FI) increased from 10 to 40. In other words, the means and SDs estimated by FIPC-BMG were closer to 0 and 1, respectively, than those estimated by the other linking procedures. And, the degree of underestimation appeared much more severe when the true ability distribution was $N(.5,1.2^2)$ than when it was $N(.25,1.1^2)$.

From Table 2 and Figure 2 it can be seen that the three other linking procedures—separate calibration with linking, concurrent calibration, and PARSCALE FIPC (FIPC-PSL)—all showed good recovery results for all conditions. They had negligible differences in true mean recovery; and, for the recovery of true SDs, these three procedures produced fairly similar results that became even more similar as sample size increased. Even for sample size N=500 and target distribution of $N(.5,1.2^2)$, the average target group means and SDs across these three linking procedures only ranged between .47 and .54 and between 1.11 and 1.27, respectively.

Figures 3 and 4 (see pages 14 and 15) summarize the MSEs, SBs, and VARs obtained using the ICC criterion. In these figures, the total length of the bar represents MSE, the shaded portion represents the SB, and the clear portion represents the VAR. First, FIPC-BMG performed poorly as expected: The less similar in ability the base and target groups were, the worse the performance. The large SB values found in Figures 3(c) and 4(c) indicate that FIPC-BMG had some systematic problem compared to the other linking procedures such as FIPC with prior update conducted by FIPC-PSL. The SB values for FIPC-BMG tended to increase as the number of fixed items decreased and as the ability distribution departed from N(0,1). Second, as

the sample size increased from 500 to 2,000, the accuracy of every linking method improved. In other words, the lengths of the bars were much shorter in Figure 4 than in Figure 3. This trend was expected because the larger sample size would result in more stable parameter calibration. Third, the performance of FIPC-PSL was similar to that for separate calibration with linking and concurrent calibration in almost every condition. That is, FIPC-PSL linked item parameters successfully regardless of the distribution difference between the base and target groups. For these three procedures, no clear pattern or improvement in accuracy was found related to the change in the number of common items (#FI = 10, 20 or 40).

Figures 5 and 6 (see pages 16 and 17) present MSEs, SBs, and VARs calculated using the TCC criterion. The overall findings were similar to those found with the ICC criterion. However, the number of fixed items seems to have more impact on the linking performance as shown in Figures 5(c), 5(f), and 5(i). As the number of fixed items increased from 10 to 40, the MSEs for every linking method dramatically decreased.

FIGURE 3. Average Mean Squared Error (=Squared Bias + Variance) across Items for the ICC Criterion (N=500)

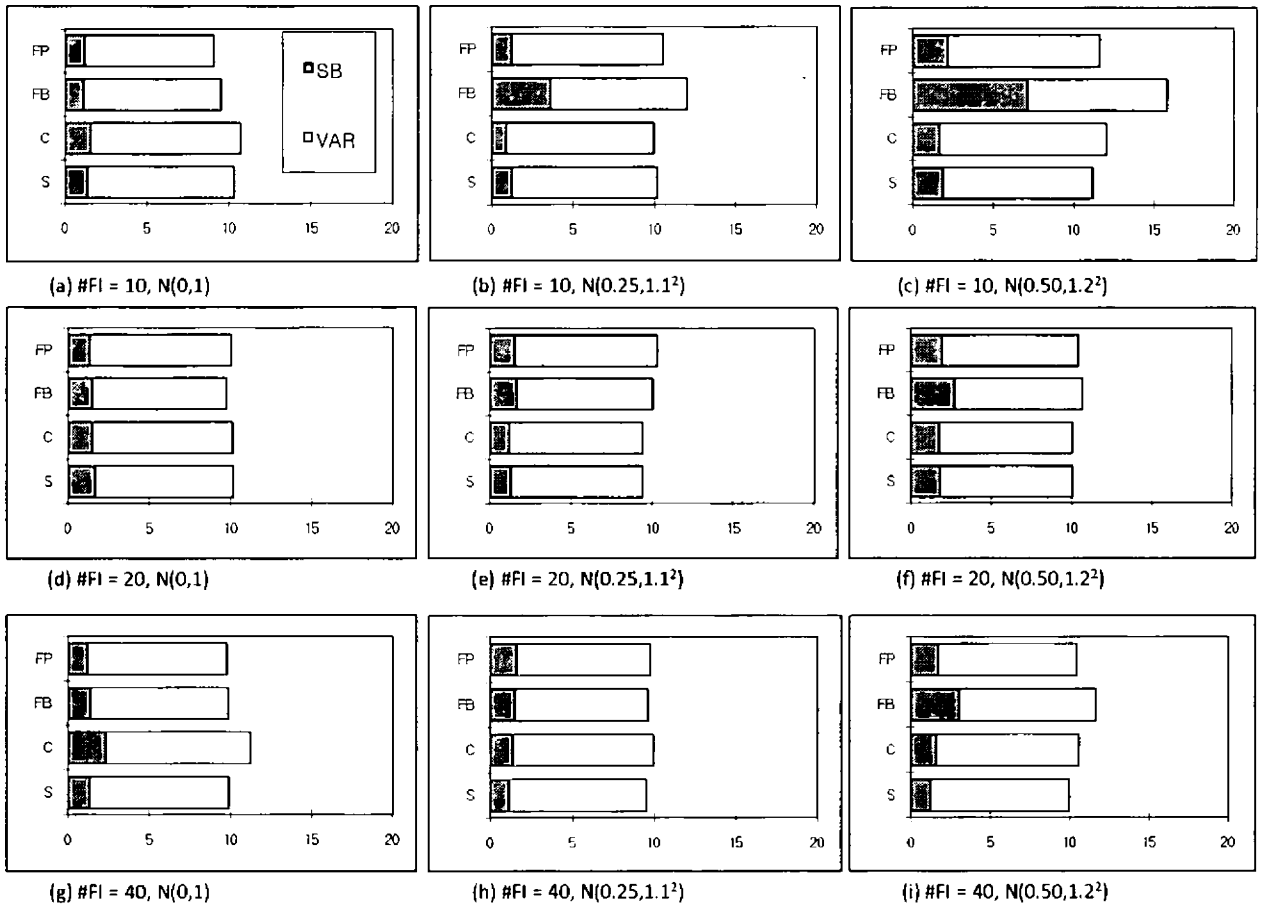


FIGURE 4. Average Mean Squared Error (=Squared Bias + Variance) across Items for the ICC Criterion (N=2,000)

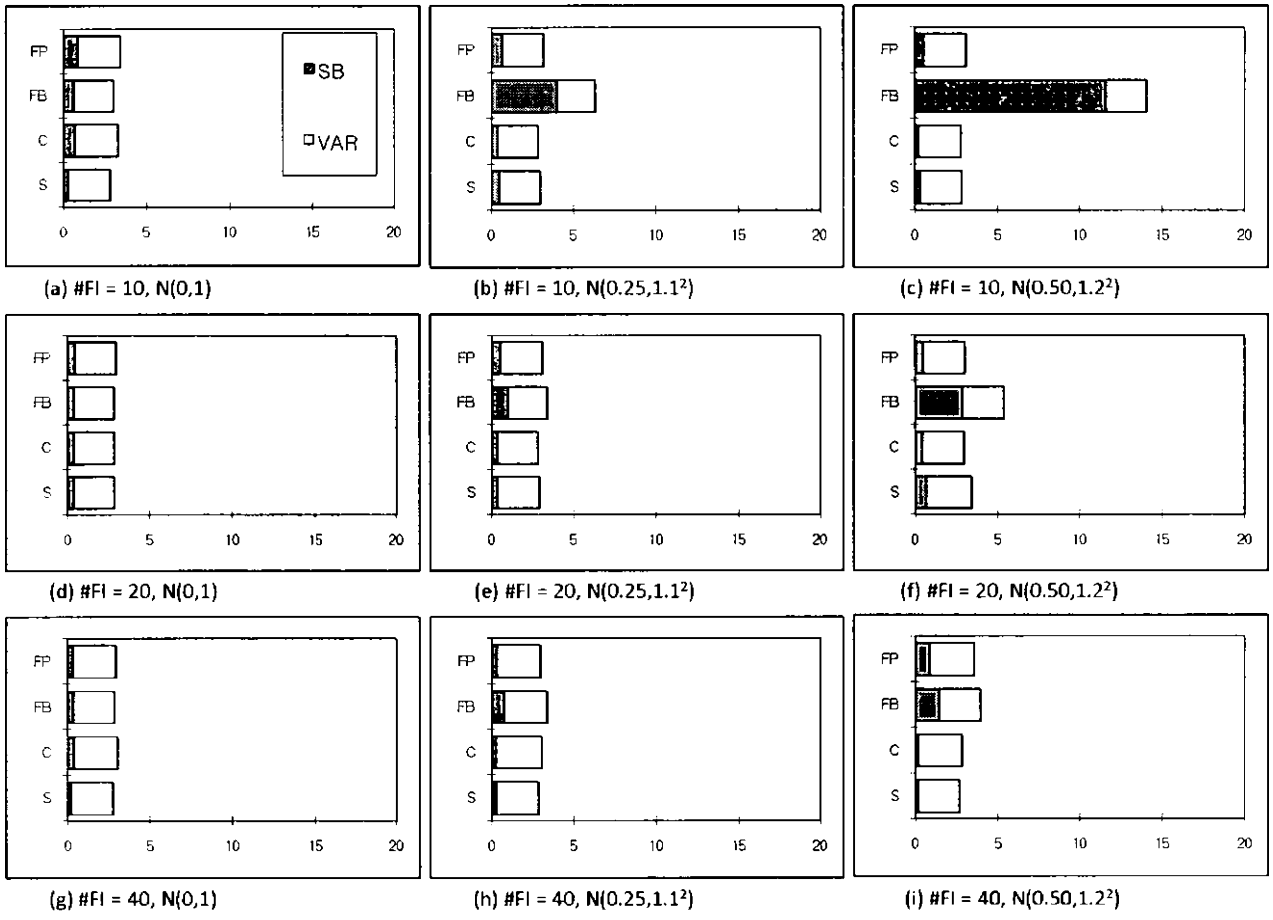


FIGURE 5. Average Mean Squared Error (=Squared Bias + Variance) across Items for the TCC Criterion (N=500)

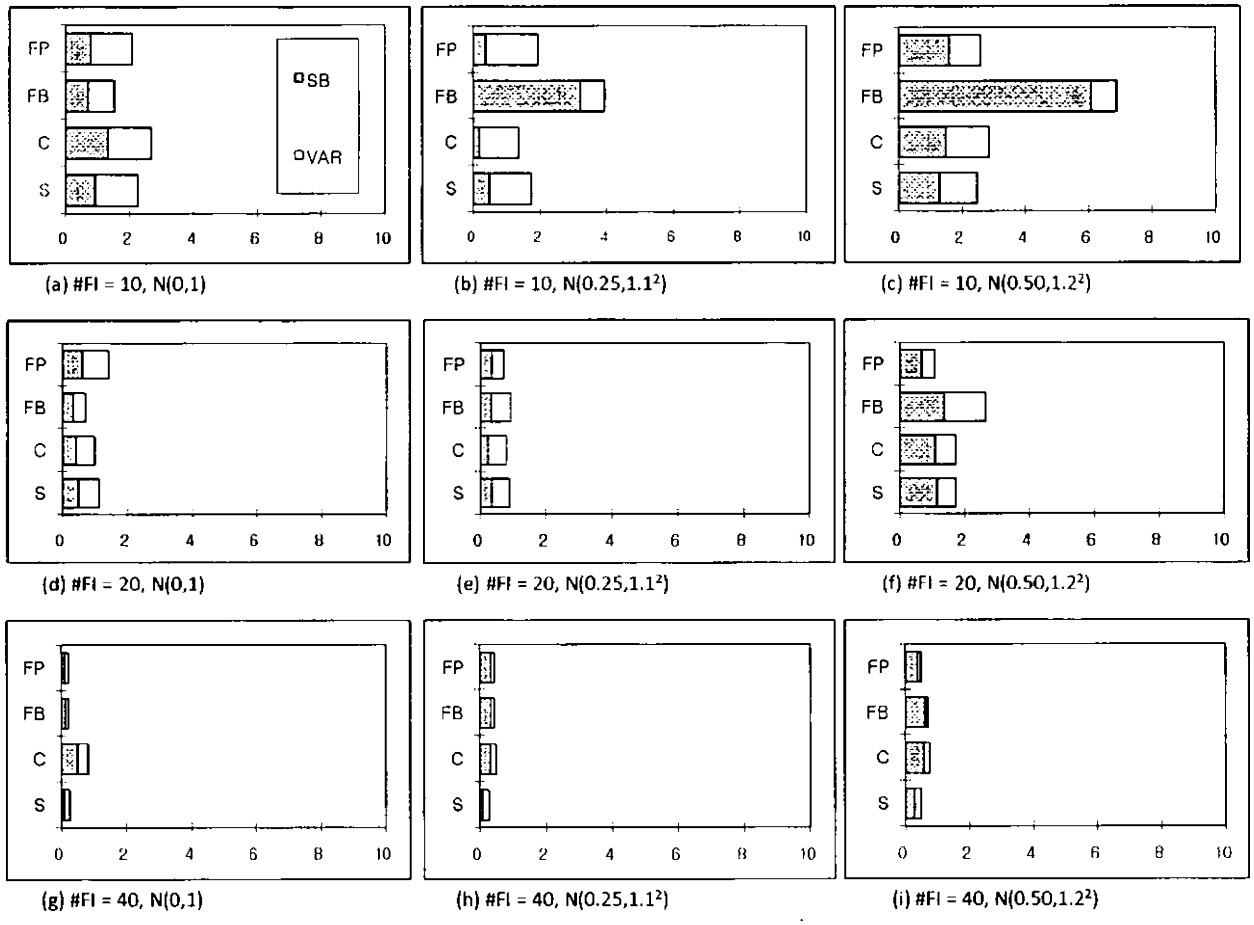
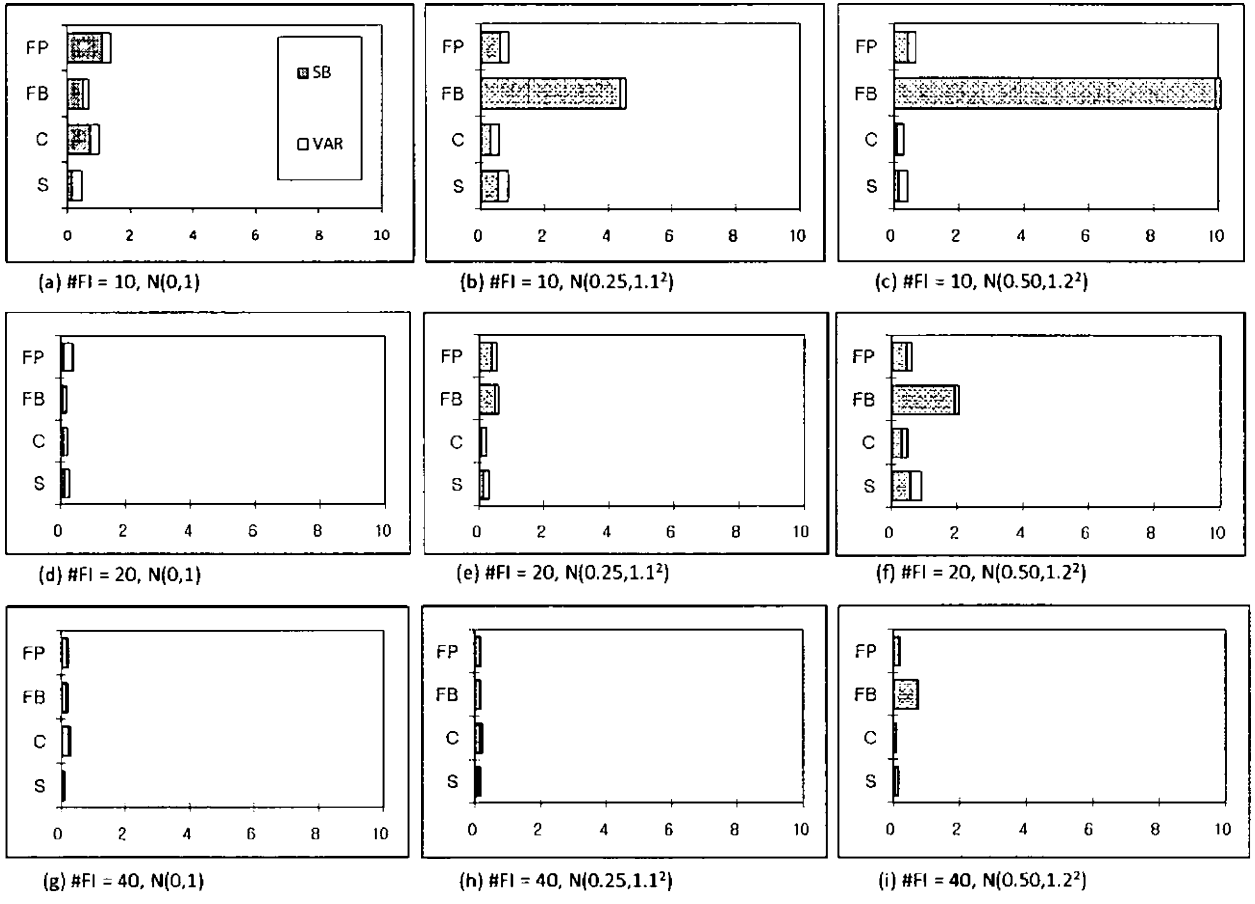


FIGURE 6. Average Mean Squared Error (=Squared Bias + Variance) across Items for the TCC Criterion (N=2,000)



Discussion and Conclusion

As expected, FIPC-BMG showed poor performance in item parameter linking when the base and target groups had nonequivalent ability distributions. When the true ability distribution of the target group was different from N(0,1), the estimated means and SDs tended to be pulled toward the mean (=0) and SD (=1) of the unchanged prior during the EM cycles. Also, for both the ICC and TCC criteria, the MSEs and SBs often were much larger than those obtained for the other linking procedures included in this study. When the base and target groups had the same ability distribution [N(0,1)], however, the MSEs from all four linking procedures were small and very similar in magnitude.

The effect of sample size for the base and target groups on the linking results was clearly observed through Figures 3 and 4 and Figures 5 and 6. When the sample size increased from 500 to 2,000, both ICC and TCC criterion values decreased remarkably. The sample size effect found in this study was consistent with that found by Hanson and Béguin (2002). In their study as the sample size increased from 1,000 to 3,000, the MSE values became smaller.

With respect to the effect of the number of common or fixed items on item parameter linking, the TCC criteria in Figures 5 and 6 showed that linking performance clearly improved at the whole test level as the number of fixed items (#FI) increased. For example, in the conditions with 2,000 examinees and a true target group ability distribution of $N(.5,1.2^2)$, the MSEs for FIPC-PSL were 0.669, 0.608, and 0.191 for #FI=10, 20, and 40, respectively. The ICC criteria in Figures 3 and 4 do not reflect clear improvement in linking performance as the number of common items increases. This result may be because the ICC criterion in this paper was calculated using only the unique (non-common) items and, thus as the number of fixed items increases, there are fewer unique items to compare. This result also might imply that use of at

least 20% common items (i.e., #FI=10 out of total 50 test items) in a test is enough for reliable linking of unique form B item parameters to the scale of form A items (see Figure 1).

This study compared the three standard methods of item calibration: concurrent calibration, separate calibration with linking, and fixed item parameter calibration (FIPC). Two different procedures for implementing FIPC were evaluated: one uses multiple EM cycles and updates the prior ability distribution multiple times during calibration and the other uses multiple EM cycles but does not update the prior ability distribution during calibration. Even though operational use of FIPC is increasing, as Paek and Young (2005) and Kim (2006) pointed out, only a few studies have compared FIPC to the concurrent and separate calibration with linking procedures that have been shown to work well in practice. Furthermore, no previous study has compared these three calibration methods while clearly identifying potential problems with FIPC and showing how to deal with these problems through the appropriate use of commercial IRT software (see Table 1 and the Appendix). Through the use of a simulation study based on actual testing program data, this paper demonstrates that concurrent calibration and separate calibration with linking produce fully acceptable results and that the correct implementation of FIPC can produce results fully comparable in accuracy to both of these procedures while other implementations of FIPC may produce severely biased results.

In summary, the FIPC procedure implemented with PARSCALE using the control cards in the Appendix works well enough to be used for contracts requiring use of FIPC. However, as this study has demonstrated, not all implementations of FIPC yield satisfactory results. Because SB and VAR in Equation 1 represent systematic and random error, respectively, the large SB values found in Figures 3, 4, 5, and 6 for FIPC-BMG indicate that this FIPC procedure failed to follow guidelines for proper application.

Additional studies could further improve our understanding of the FIPC procedure. First, future work could include more commercial IRT software with options available for fixing item parameters such as MULTILOG (Thissen, 1991) and WINSTEPS (Linacre, 2003). Second, the performance of these linking procedures needs to be compared when the set of common items consists of either polytomous items or mixed-format items. Finally, more conditions and simulation factors could be considered. For example, it would be interesting to include other true target ability distributions such as a skewed one. Also, inclusion of some misfitting items among the common or unique items could be considered to make the generated data sets even more realistic. And, a simulation study could be conducted in the context of a computerized adaptive testing program.

The results of this study are important for practitioners as it is demonstrated that some implementations of FIPC will yield acceptable calibration results comparable to those obtained with concurrent and separate calibration with linking and that other implementations will not yield acceptable calibration results. Knowing how to properly implement FIPC is very important in practice as contracts may require use of FIPC and it can be less time consuming to use FIPC than other calibration procedures.

References

Baldwin, S. G., Baldwin, P., & Nering, M. L. (2007). *A comparison of IRT equating methods on recovering item parameters and growth in mixed-format tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24.

Keller, R. R., Keller, L. A., & Baldwin, S. (2007). *The effect of changing equating methods on monitoring growth in mixed-format tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 355-381.

Linacre, J. M. (2003). *WINSTEPS* [Computer Program]. Chicago: MESA Press.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

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

Paek, I., & Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education, 18*, 199-215.

Skorupski, W. P., Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). *An evaluation of item response theory equating procedures for capturing growth with tests composed of dichotomously scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Thissen, D. (1991). *Multilog user's guide: Multiple categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software International.

Appendix

1. Concurrent Calibration for both base and target groups (BILOG-MG)

```
>COMMENT
>GLOBAL DFNAME='BLMcom200.dat',NPARM=3,SAVE;
>SAVE  PARM='BLMcom200.par';
>LENGTH NITEMS=60;
>INPUT  NTOT=60, NID=4, NGROUP=2, NFNAME='c:\FIPC\simu\keynot.txt';
>ITEMS  INUM=(1(1)60),INAMES=(OL01(1)OL10, CO01(1)CO40, NE01(1)NE10);
>TEST  TNAME=Math;
>GROUP1 GNANE='BASE', LENGTH=50, INUM=(1(1)50);
>GROUP2 GNANE='TARGET', LENGTH=50, INUM=(21(1)60);
(4A1,1X,I1,1X,60A1)
>CALIB NQPT=11, cycles=3000, CRIT=0.001, REF=1, TPRIOR;
>SCORE  NOPRINTS;
```

2. Separate Calibration for a target group (BILOG-MG)

```
>COMMENT
>GLOBAL DFNAME='new200.dat',NPARM=3,SAVE;
>SAVE  PARM='BLMnew200.par';
>LENGTH NITEMS=50;
>INPUT  NTOT=50, NALT=5, NID=4;
>ITEMS  INUM=(1(1)50),INAMES=(CO01(1)CO40, NE01(1)NE10);
>TEST  TNAME=Simulation;
(4A1,T1,50A1)
>CALIB  NQPT=11, cycles=3000, CRIT=0.001, TPRIOR;
>SCORE  NOPRINTS;
```

3. Fixed Item Parameter Calibration for a target group (BILOG-MG)

```
>COMMENT
>GLOBAL DFNAME='new200.dat', PRNAME='BLMOLD200.PRM', NPARM=3, SAVE;
>SAVE  PARM='BLMfix200.par';
>LENGTH NITEMS=50;
>INPUT  NTOT=50, NALT=5, NID=4;
>ITEMS  INUM=(1(1)50),INAMES=(CO01(1)CO40, NE01(1)NE10);
>TEST  TNAME=Math, FIX=(1(0)40,0(0)10);
(4A1,T1,50A1)
>CALIB  NQPT=11, cycles=3000, CRIT=0.001, TPRIOR, NOADJUST;
>SCORE  NOPRINTS;
```

4. Fixed Item Parameter Calibration for a target group (PARSCALE)

```
>COMMENT
>FILE   DFNAME='new200.dat', IFNAME='PSLold200.prm', SAVE;
>SAVE   PARM='fix200.par';
>INPUT  NIDCH=4, NTOTAL=50, NTEST=1, LENGTH=50, NFMT=1;
(4A1, T1, 50A1)
>TEST   TNAME=Math, ITEM=(01(1)50), NBLOCK=50,
     INAMES=(
     CO01, CO02, CO03, CO04, CO05, CO06, CO07, CO08, CO09, CO10,
     CO11, CO12, CO13, CO14, CO15, CO16, CO17, CO18, CO19, CO20,
     CO21, CO22, CO23, CO24, CO25, CO26, CO27, CO28, CO29, CO30,
     CO31, CO32, CO33, CO34, CO35, CO36, CO37, CO38, CO39, CO40,
     NE01, NE02, NE03, NE04, NE05, NE06, NE07, NE08, NE09, NE10);
>BLOCK1  BNAME=COMMON, NITEM=1, NCAT=2,
     ORI=(0,1), MOD=(1,2), GPARM=0.2, GUESS=(2, EST), REP=40, SKIP;
>BLOCK2  BNAME=UNIQUE, NITEM=1, NCAT=2,
     ORI=(0,1), MOD=(1,2), GPARM=0.2, GUESS=(2, EST), REP=10;
>CALIB  PARTIAL, LOGISTIC, SCALE=1.7, NQPT=41, CYCLE=(3000,1,1,1,1),
     FREE=(NOADJUST, NOADJUST), POSTERIOR, NEWTON=0, CRIT=0.001, ITEMFIT=10, SPRIOR, GPRIOR;
>SCORE  ;
```