Asia Pacific Education Review
2008, Vol. 9, No.3, 248-261.

A New Approach to Test Score Equating Using Item Response Theory with Fixed C-Parameters

Guemin Lee
Yonsei University
Korea

Anne R. Fitzpatrick
Educational Testing Services
USA

Because parameter estimates from different calibration runs under the IRT model are linearly related, a linear equation can convert IRT parameter estimates onto another scale metric without changing the probability of a correct response (Kolen & Brennan, 1995, 2004). This study was designed to explore a new approach to finding a linear equation by fixing C-parameters for anchor items in IRT equating. A rationale for fixing C-parameters for anchor items in IRT equating can be established from the fact that the C-parameters are *not* affected by any linear transformation. This new approach can avoid the difficulty in getting accurate C-parameters for anchor items embedded in the application of the IRT model. Based upon our findings in this study, we would recommend using the new approach to fix C-parameters for anchor items in IRT equating.

Key words: Equating, Item Response Theory, Stocking and Lord Method

Introduction

When a new test is administered as part of a high stakes testing program, the test must usually be linked to a previously established score scale. To do this, it is common to administer and calibrate a set of anchor items with the new test. The anchor items have parameter estimates on the previously established score scale. The new parameter estimates for these anchor items are used to link the new test to the old scale to establish score equivalents between old and new forms. This process is usually called IRT equating.

When the IRT model holds, the parameter estimates from different calibration runs are linearly related. A linear equation can convert IRT parameter estimates onto anther scale metric without changing the probability of a correct response in IRT models. The equating constants, *m1* (slope) and *m2* (intercept), are used to transform the new parameter estimates so that they are expressed on the old score scale using the following equations,

$$\begin{cases} B* = m1B + m2 \\ A* = A / m1 \\ C* = C \end{cases} \tag{1}$$

where *A*, *B*, and *C* are the item parameters in the 3PL model, and the * mark represents transformed parameters. Because C-parameters are on the probability metric, those remain the same before and after transformation.

There are several ways of determining equating

Guemin Lee, Department of Education, Yonsei University, Korea; Anne R. Fitzpatrick, Educational Testing Services, USA.

This work was supported by a Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-003-B00229).

Correspondence concerning this article should be addressed to Guemin Lee, Department of Education, Yonsei University, 134 Shinchon-Dong, Seodaemoon-Ku, Seoul, Korea. E-mail: guemin @yonsei.ac.kr

constants, *m1* and *m2*. Probably the most straightforward way would be to substitute the means and standard deviations of the item parameter estimates in equation 1. Marco (1977) described the mean/sigma method, in which the means and standard deviations of the b-parameter estimates from the common items are used to find *m1* and *m2*. Loyd and Hoover (1980) developed the mean/mean method. In this method, the mean of the a-parameter estimates is used to find the *m1* constant and the mean of the b-parameter estimates is used to find the *m2* constant. However, the mean/sigma and mean/mean methods have been criticized in that the equating constants could be overly influenced by the differences of item parameter estimates (Kolen & Brenna, 1994).

Characteristic curve methods were proposed to consider all of the item parameter estimates simultaneously in finding *m1* and *m2* equating constants. Haebara (1980) developed a function to express the difference between *item characteristic curves* by summing up the squared differences between the old and new form items. In contrast to the Haebara approach, Stocking and Lord (1983) used the sum of squared differences over items. That is, the summation is taken over items for each set of parameter estimates before squaring. Thus, Stodking and Lord's method can be referred to as the *test characteristic curve* method rather than item characteristic curve method.

Several studies have been conducted to compare the relative appropriateness of each method in determining equating constants, m1 and m2 (Baker & Al-Kari, 1991; Hanson & Beguin, 2002; Hung, Wu, & Chen, 1991; Kim & Cohen, 1992; Ogsawara, 2001; Way & Tang, 1991). The general conclusion from these studies is that the characteristic curve methods produce more accurate results. Consequently, the characteristic curve methods have been adopted by many testing programs to determine equating constants in finding scale equivalents under IRT models.

The Stocking and Lord (SL) characteristic curve method (Stocking & Lord, 1983) is one of the most widely used equating procedures based on item response theory (IRT) methodology. This study focused on a new approach to test score equating with the SL method, but the rationale of implementing the new approach can be applied to other equating methods. Described within the context of linking tests, the SL procedure involves finding the linear transformation that minimizes the function

$$F = \frac{1}{N}\sum_{a=1}^{N}(\hat{\tau}_a - \hat{\tau}_a^*)^2 \quad , \qquad (2)$$

where N is the number of examinees in the simulated group, and $\hat{\tau}_a$ and $\hat{\tau}_a^*$ refer to the true scores obtained by examinees when scored using, respectively, the original item parameters and the new parameter estimates for the anchor items.

The three-parameter logistic (3PL) model is commonly used to scale multiple-choice (MC) items. The 3PL model defines the probability that an examinee with ability θ will correctly answer the *j*th item as

$$P_j(\theta) = P(X_j = 1|\theta) = C_j + \frac{1-C_j}{1+\exp\{-1.7A_j(\theta - B_j)\}} \ , \qquad (3)$$

where A_j, B_j, and C_j, respectively, refer to the discrimination, location, and "pseudo-guessing level" parameters of the item (Lord, 1980).

Previous studies have indicated that the C-parameter in the 3PL model is unstable. For example, Wingersky, Barton, and Lord (1982) showed that C-parameters could not be accurately estimated for very easy and moderately easy items with low degrees of discrimination. This finding makes it possible to draw some implications in the IRT equating context. Since examinees tend to improve their test performance over many years of a testing program, anchor items used in later versions of a test often become easier and less discriminating than they were when first administered. The relative easiness and/or lower discrimination levels of these items can make it difficult to obtain accurate C-parameter estimates (Lord, 1980; Thissen & Wainer, 1982). Poor estimation of the C-parameters, in turn, can diminish the accuracy of the estimates for the A and B parameters (Thissen & Wainer, 1982).

The main purpose of this study was to explore a new approach to fixing C-parameters for anchor items to be used in the SL equating. A rationale for fixing C-parameters for anchor items in SL equating could be established from the fact that the C-parameters are *not* affected by any linear transformation. That is, the C-parameters remain constant before and after equating. This new approach can avoid a difficulty in getting accurate C-parameters for anchor items embedded in the applications of the 3PL model into SL equating. It is expected that the logic of the new approach will be applicable to other IRT-based equating methods as well.

Method

Data Sources

The data were collected during the standardization of reading and mathematics tests of the TerraNova, a national standardized achievement test batteries (CTB/McGraw-Hill, 2001). Students in Grades 4, 7, and 11 were tested. The test in each subject area and grade was calibrated using about 4,000 to 5,000 students per grade. About 1,500 to 2,000 of these students took both the new test and the corresponding test in an older edition of the achievement battery. General raw score descriptive statistics for data used in this study are presented in Table 1.

Equating Procedures

Equating procedures in the 1999 standardization for the achievement tests involved two stages. Two different equating designs were implemented, *a single group with counterbalancing* design in one stage and *a common items with nonequivalent groups design* in anther stage (Kolen & Brennan, 1995).

In the first stage, the item parameters for the old and

Table 1
Descriptive Statistics for Data Sources Used in This Study

	Sample Size	No. of Total Items	No. of Anchor Items	Mean	Standard Deviation	KR20
Grade 4						
Reading						
New Form	4,500	35		22.5	6.91	0.88
Old Form	1,783	35	35	23.9	7.51	0.91
Mathematics						
New Form	4,628	32		19.4	6.12	0.84
Old Form	1,884	32	32	20.8	6.36	0.88
Grade 7						
Reading						
New Form	3,869	33		20.1	6.94	0.88
Old Form	1,577	33	33	20.0	7.25	0.91
Mathematics						
New Form	4,118	32		18.8	6.90	0.88
Old Form	1,705	32	32	18.6	6.40	0.89
Grade 11						
Reading						
New Form	4,424	34		17.7	7.36	0.88
Old Form	2,038	34	34	18.6	7.44	0.92
Mathematics						
New Form	4,929	25		12.5	5.21	0.82
Old Form	2,215	25	25	12.5	5.22	0.89

Note. Because a single group with counterbalancing design was implemented, all test items in the old form served as anchor items in this study.

new tests were *concurrently* calibrated. This concurrent calibration places item parameters for the two tests on the same scale (Hambleton, Swaminathan, & Rogers, 1991; Hambleton, 1989; Cook & Eignor, 1991).

Item parameter estimates of the new test are transformed onto the scale of the old test items when those were standardized five years before. This second stage of equating can be accomplished by applying equating constants that derived from using the SL or other IRT-based equating methods. Due to the fact that the item parameter estimates from both the current and target year standardization for the old test (anchor items) are used, the equating design for the second stage is the design involving common items and nonequivalent groups (Kolen & Brennan, 1995).

Procedures for Fixing C's

Items in the new and old test forms in each grade and content were calibrated concurrently using two different approaches. As one approach, the C-parameter estimates for all items were estimated in the usual way. In another approach, the C-parameters for the items of anchor test were set equal to the values obtained when old test was standardized five years before.

The procedures for fixing the C-parameters are related to the estimation algorithm for item parameters. In this study, the marginal-maximum-likelihood (MML) estimation method with EM cycles was investigated. To fix C-parameters in this estimation algorithm, inputs for C-parameters for starting values and for every E-step were set to the values from the standardization five years before. However, the A- and B-parameters were free to be estimated. Interested readers regarding detailed discussion about the MML and EM could consult Bock and Aitkin (1981) and Yen (1990).

Analyses

Two computer application programs were used for concurrent calibrations, PARDUXMJ (Bucket, 2000) that fixed C-parameters for anchor items and PARDUXMX (Bucket, 1996) without fixing C-parameters for anchors. The general characteristics for these two calibration approaches were compared in terms of the number of EM cycles, model-data fit results, the number of non-converging items, and so forth. The item parameter estimates from the two calibration methods before equating were plotted together for new and old test forms separately.

The SL procedure was then implemented to link the new and old tests. To evaluate the equating functions which resulted from the two different calibrations, the Stocking and Lord's F-values were examined. The smaller F-value is likely to represent the better equating. The graphical comparisons of test characteristic curves for anchor items with equated item parameters from two methods were compared using TCC of five-year-before standardization as a target. The following statistics were also computed as a means of evaluating equating functions,

$$Diff. = \bar{\alpha}_{new} - \bar{\alpha}_{old} \tag{4}$$

$$RMSD = \sqrt{\frac{\sum_{i=1}^{I} (\alpha_{i,new} - \alpha_{i,old})^2}{I}} \tag{5}$$

$$Ratio = \frac{SD(\alpha_{new})}{SD(\alpha_{old})} \tag{6}$$

$$r = corr.(\alpha_{new}, \alpha_{old}) \tag{7}$$

where α_{new} and α_{old} represent a vector of estimates for each of A-, B-, and C-parameters for anchor items, and *I* is the number of anchor items.

To investigate the effects of the two different approaches, the same set of students who responded to old test form (anchor item set) was scored three times using three different sets of item parameter estimates. One was made up of the original item parameter set that was calibrated five years before. The other two are the equated item parameter sets from fixing and without fixing C-parameters for these anchor items. The student scale scores scored from applying original item parameter set served targets for comparing the two methods.

The mean, median, standard deviation of the student scale scores was compared. The percentages of students at lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) were also examined. The scale scores corresponding to selected percentiles were presented and compared in relation to the two methods relative to the values of the target.

Results

Concurrent Calibration

Table 2 shows the general characteristics for the calibration results with fixing C-parameters (Fix_C) and without fixing C-parameters (w/o Fix_C) for the anchor items. In some grades and in terms of some of the content, the Fix_C calibration converged with fewer EM cycles, but in other grades and in terms of other content, it required more EM cycles than did the w/o Fix_C calibration. In general, both Fix_C and w/o Fix_C calibrations needed a similar number of EM cycles. There were no non-converging items and one or two poor-fit items in both calibrations. Because tests used in this study are standardized operational tests that are composed of carefully selected items among tryout items, it may not be surprising to find a few faulty items in them. Additionally, similar numbers of items having a maximum A value were reported in both calibrations. The maximum A was set to a certain value and assigned to items that have estimates for A-parameter greater than this value. Default C was set to 1/(no. of choices+1) and the number of items with default C was counted for only non-anchor items. According to the calibration results, it seems reasonable to conclude that both the Fix_C and w/o Fix_C procedures had similar calibration characteristics.

Table 2

Summary Statistics for Item Calibration Results With and Without Fixing C-Parameters for Anchor Items

	No. of EM Cycles	Non-Converging Items	Poor-fit Items	No. of Maximum A	No. of Default C
Grade 4					
Reading					
Fix_C	36	0	1	0	10
w/o Fix_C	49	0	0	0	10
Mathematics					
Fix_C	32	0	0	0	5
w/o Fix_C	22	0	0	0	5
Grade 7					
Reading					
Fix_C	42	0	1	1	4
w/o Fix_C	36	0	1	1	4
Mathematics					
Fix_C	42	0	0	0	0
w/o Fix_C	50	0	0	0	0
Grade 11					
Reading					
Fix_C	50	0	1	2	2
w/o Fix_C	50	0	1	1	2
Mathematics					
Fix_C	37	0	2	0	2
w/o Fix_C	50	0	2	1	1

Note. Fix_C = fixing C-parameters for anchor items; w/o Fix_C = without fixing C-parameters for anchor items.

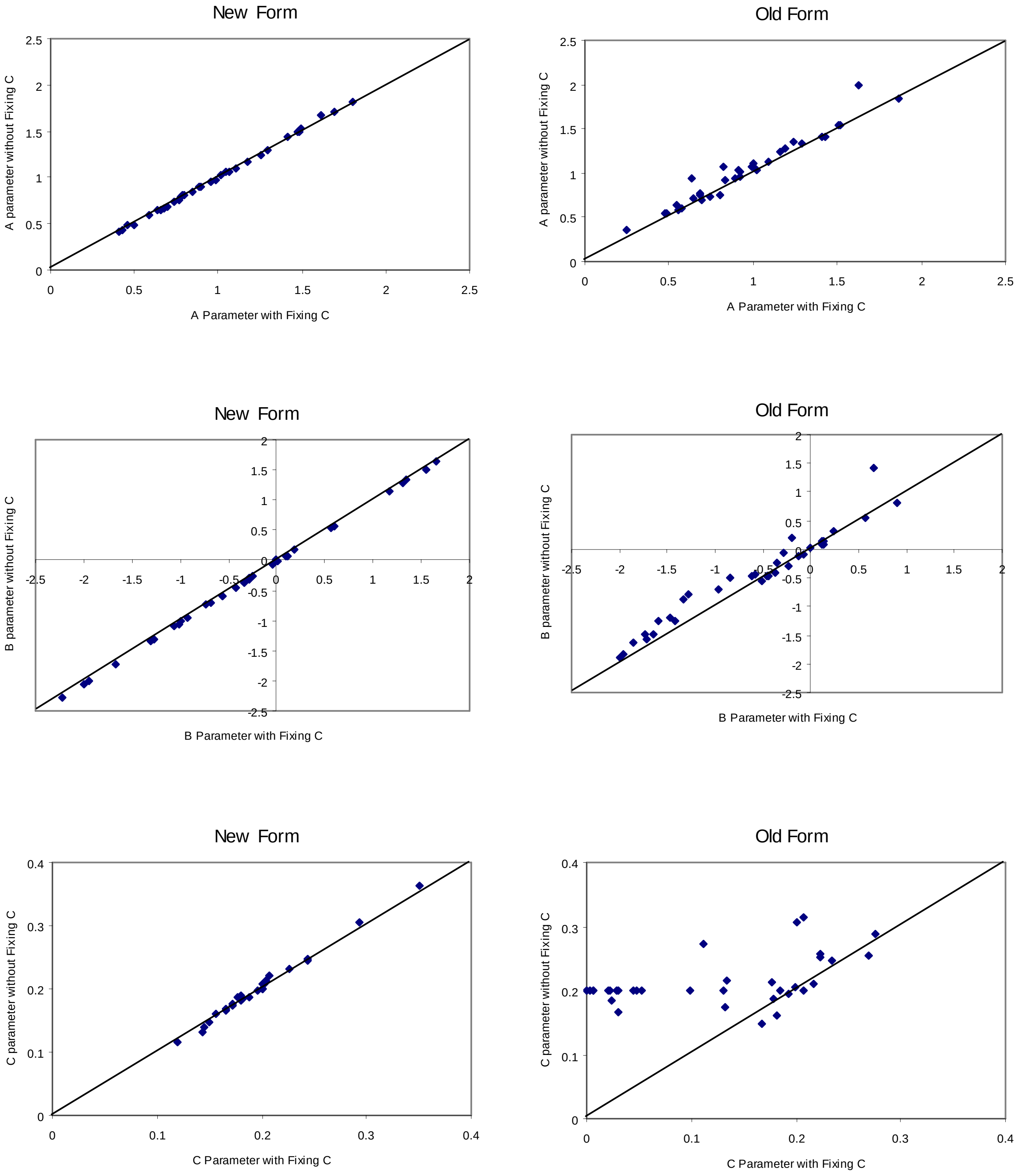

Figure 1. Comparison of item parameter estimates from the calibration with and without fixing C-parameter for anchor items in grade 4 Reading test.

As one method of comparing the Fix_C and w/o Fix_C calibration results, item parameter estimates from two calibrations before equating are plotted together. Figure 1 shows these for grade 4 Reading Comprehension test. Similar plots and trends were found in other grades content areas that are not presented in current paper.

We can observe well-aligned estimates for all of A, B, and C parameters in the new test form (the left-hand side three plots). This means that the calibration procedures were not significantly affected by the Fix_C and w/o Fix_C specifications in estimating item parameters for non-anchor items. In contrast, relatively greater differences were found in item parameter estimates of the old form items (the right-hand side three plots), especially in terms of C-parameters.

The C-parameters for anchor items were fixed and set to values of five-year-before standardization in the Fix_C calibration. Thus, plots of C-parameter estimates for anchor items can be considered as showing the comparison of targets and their corresponding estimates. The results of this study have showed some degree of instability of C-parameter estimates and this is something which is consistently indicated by previous studies. The instability of C-parameter estimates can be related to the inaccuracy of A and/or B parameter estimates in the way that Thissen and Wainer (1982) have pointed out.

Stocking and Lord Equating

Table 3 provides equating constants, M1 and M2, and Stocking and Lord's F-values after SL equating by using

Table 3

Equating Constants, M1 and M2, and Stocking and Lord's F-Values Using Item Parameter Sets from Fixing and Without Fixing C-Parameters for Anchor Items

	Equating Constants		
	M1	M2	F-Value
Grade 4			
Reading			
w Fix_C	33.89	633.14	0.43
w/o Fix_C	33.12	633.45	1.16
Mathematics			
w Fix_C	33.31	616.78	0.09
w/o Fix_C	33.16	614.67	0.17
Grade 7			
Reading			
w Fix_C	37.17	656.45	0.12
w/o Fix_C	37.22	656.00	0.11
Mathematics			
w Fix_C	36.79	663.48	0.13
w/o Fix_C	36.28	665.15	0.14
Grade 11			
Reading			
w Fix_C	37.93	687.28	0.22
w/o Fix_C	37.23	692.58	0.45
Mathematics			
w Fix_C	41.72	716.87	0.19
w/o Fix_C	42.75	715.11	0.38

Note. Fix_C = fixing C-parameters for anchor items; w/o Fix_C = without fixing C-parameters for anchor items.

item parameter estimates from two calibrations with Fix_C and w/o Fix_C specifications. The F-value can be conceived as a criterion to evaluate the SL equating because the SL equating finds equating constants such that those minimize this F. Thus, the smaller F-value implies a better equating function. The Fix_C method produced smaller F-vales than did the w/o Fix_C method in grade 4 Reading Comprehension and grade 11 Reading Comprehension and Mathematics tests. For grade 4 Mathematics and grade 7 Reading Comprehension and Mathematics, both the Fix_C and w/o Fix_C had similar F-values. Thus, it seemed reasonable to conclude that the Fix_C procedure produced the better equating functions in grade 4 Reading Comprehension and grade 11 Reading Comprehension and Mathematics.

The SL procedure is one of the IRT-based equating methods that utilize test characteristic curves (TCC's). Thus, investigating TCC's would appear to be a meaningful exercise. Three TCC's are presented in Figure 2 for the grade 4 Reading Comprehension test. The TCC with item parameters calibrated at five-years-before standardization served as a target TCC. Next, the TCC's formulated by item parameter sets from the Fix_C and w/o Fix_C procedures were compared. The TCC closer to the target indicates the better equating.

The TCC of the Fix_C procedure was aligned better to the target than was that of the w/o Fix_C procedure. A non-negligible difference between TCC's of the w/o Fix_C procedures and the target were found in lower ability scale. This is consistent with the previous results that presented the large F-value for the w/o Fix_C procedure. The F-value for the w/o Fix_C was 1.16, but it was just 0.43 for the Fix_C. Based upon Figure 2 and Table 3, the relative large F-value for the w/o Fix_C in grade 4 Reading Comprehension test seemed mainly caused by the difference between TCC's in the lower ability scale. Moreover, the difference between TCC's in lower ability scales is likely to be related to the poor estimation of C-parameters because the lower-asymptote of TCC is the sum of lower-asymptotes of items. Similar trends were found in grade 11 Reading Comprehension and Mathematics tests.

Figure 3 provides three TCC's for grade 4 Mathematics tests. All three TCC's were very similar, though the target and the Fix_C were a little bit closer in terms of the lower

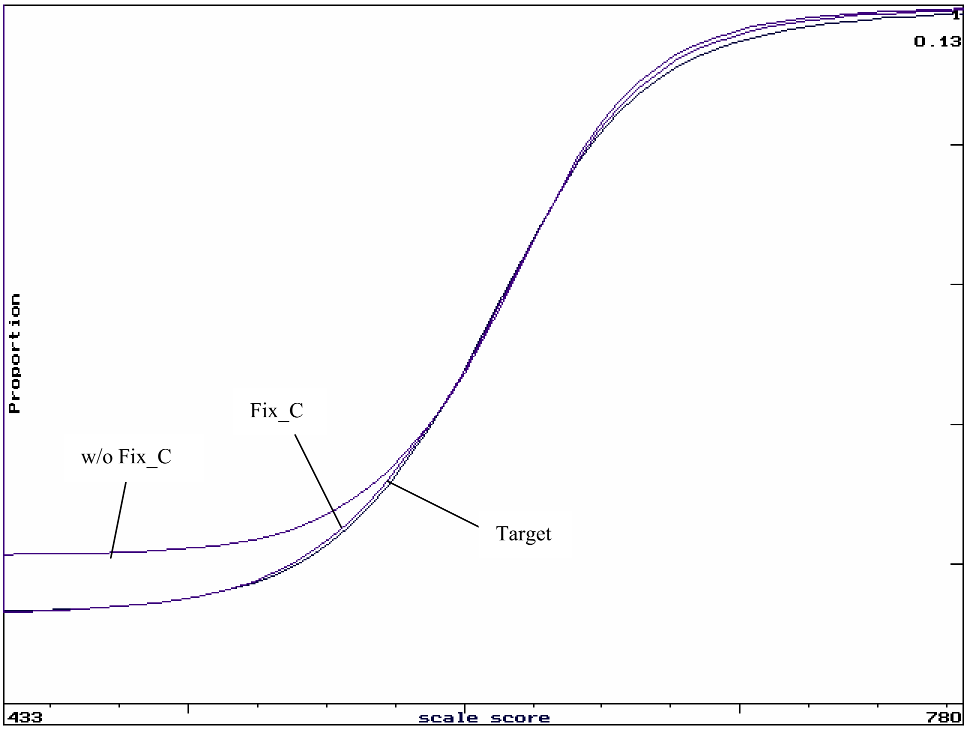

Figure 2. Test characteristic curves formulated by item parameter sets from the target and the fixing (Fix_C) and without fixing c-parameters (w/o Fix_C) for anchor items of grade 4 Reading test.

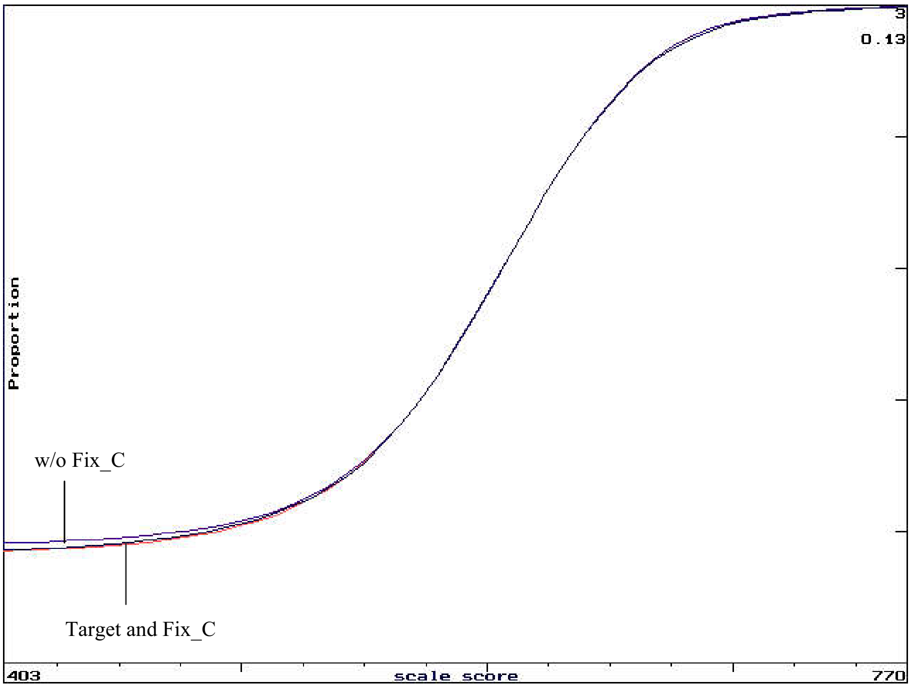

Figure 3. Test characteristic curves formulated by item parameter sets from the target and the fixing (Fix_C) and without fixing c-parameters (w/o Fix_C) for anchor items of grade 4 Mathematics test.

ability scale. However, this relative closeness seemed sufficiently small so as to be ignored. These trends were also observed in grade 7 Reading Comprehension and Mathematics tests.

Several summary statistics to evaluate equated item parameters are presented in Table 4 for both the Fix_C and w/o Fix_C procedures. These statistics were computed with equations derived in the previous section for only anchor items using item parameters from the five-year-before standardization as a target.

First of all, it should be mentioned that there was zero DIFF. and RMSD for the Fix_C procedure because it set C-parameters to the target values. Thus, it is not meaningful to compare C-parameter related statistics of the Fix_C and w/o Fix_C procedures. Only relative comparisons among magnitudes within the w/o Fix_C category will give any meaningful information. Relatively large DIFF. and RMSD values were reported for grade 4 Reading and grade 11 Reading and Mathematics tests in the comparison of C-parameters. The grades and content areas that showed relatively large DIFF. and RMSD are consistent with those that the investigation of F-values and TCC's indicated. It was difficult to make meaningful interpretations for other indices. In other words, Diff. and/or RMSD rather than Ratio and/or r could be considered indices. This could be recognized as evidence of possible problems in IRT equating.

Scale Score Comparison

Table 5 shows means, standard deviations, medians, and percentages of students at LOSS and HOSS using different sets of item parameter estimates from fixing and without fixing C-parameters for anchor items. The statistics for student scale scores computed from using item parameters of five-year-before standardization served as criteria to evaluate the performance of equated item parameters from two different calibrations and equating functions.

For a grade 4 Reading Comprehension test, the mean and standard deviation of student scale scores from the Fix_C procedure were more similar to those of the target than were those of the w/o Fix_C procedure. For example,

Table 4

The Comparison of Equated Item Parameter Estimates of With and Without Fixing C-parameters for Anchor Items

	A-Parameter Comparison				B-Parameter Comparison				C-Parameter Comparison			
	Diff.	RMSD	Ratio	r	Diff.	RMSD	Ratio	r	Diff.	RMSD	Ratio	r
Grade 4												
Reading												
w Fix_C	0.002	0.004	1.074	0.932	-1.837	8.698	0.984	0.951	0.000	0.000	1.000	1.000
w/o Fix_C	0.005	0.006	1.119	0.954	3.702	8.805	0.918	0.958	0.086	0.115	0.433	0.463
Mathematics												
w Fix_C	0.000	0.003	0.956	0.949	0.375	5.443	0.970	0.988	0.000	0.000	1.000	1.000
w/o Fix_C	0.001	0.003	0.985	0.936	1.349	7.795	0.958	0.975	0.012	0.072	0.900	0.435
Grade 7												
Reading												
w Fix_C	0.001	0.004	1.048	0.941	0.255	4.086	1.028	0.986	0.000	0.000	1.000	1.000
w/o Fix_C	0.001	0.005	0.997	0.900	1.452	7.004	1.041	0.961	0.016	0.052	0.642	0.702
Mathematics												
w Fix_C	0.000	0.005	1.332	0.753	0.860	7.222	0.892	0.984	0.000	0.000	1.000	1.000
w/o Fix_C	0.000	0.006	1.178	0.668	1.832	7.962	0.965	0.976	-0.004	0.076	0.919	0.366
Grade 11												
Reading												
w Fix_C	0.000	0.005	0.973	0.855	-0.673	3.728	0.974	0.994	0.000	0.000	1.000	1.000
w/o Fix_C	0.001	0.005	0.911	0.858	1.801	6.937	0.945	0.979	0.036	0.070	0.546	0.340
Mathematics												
w Fix_C	0.000	0.002	1.032	0.967	-0.599	6.078	1.034	0.983	0.000	0.000	1.000	1.000
w/o Fix_C	0.001	0.004	0.938	0.934	4.652	16.379	1.001	0.875	0.044	0.080	1.111	0.491

Note. Fix_C = fixing C-parameters for anchor items; w/o Fix_C = without fixing C-parameters for anchor items.

the mean of the w/o Fix_C procedure is 7.2 scale score points lower than that of the target, but the Fix_C is only 1.8 points lower. The standard deviation of the w/o Fix_C was much larger than that of the target, while that of the Fix_C was very similar to that of the target. Similar trends were observed in grade 11 Reading Comprehension and Mathematics tests though the magnitudes of differences varied across grades and content areas. For grade 4 Mathematics and grade 7 Reading Comprehension and Mathematics tests, both the Fix_C and w/o Fix_C and the target produced similar means and standard deviations.

The comparison of student scale scores lead to the consistent interpretations to which the evaluation of equating functions reached. That is, the larger F-values, more different TCC's, and more discrepant C-parameters were reported for the w/o Fix_C procedure in grade 4 Reading Comprehension and grade 11 Reading Comprehension and Mathematics tests. For the same grades and content areas, the larger differences of means and standard deviations were found for the w/o Fix_C procedure than for the Fix_C procedure compared to the target.

Both the Fix_C and w/o Fix_C procedures and the target produced almost the same medians across grades and content areas. This means that it is likely to get similar scale scores when applying any among three item parameter sets if student ability scores are in the middle range of the score scale. In contrast, somewhat different LOSS information for the w/o Fix_C was reported in grade 4 Reading Comprehension and grade 11 Reading Comprehension and Mathematics tests. The w/o Fix_C procedure produced a relatively large percentage of students at LOSS than did the Fix_C and target. However, exactly the same amount of HOSS information was obtained for the two procedures and the target.

Table 5
Mean, Median, Standard Deviation of Student Scale Scores and Percentages of Students at LOSS and HOSS Using Items Parameter Sets from Fixing and Without Fixing C-Parameters for Anchor Items

	Sample Size	Mean	Standard Deviation	Median	% at LOSS	% at HOSS
Grade 4						
Reading						
Target	2123	629.3	48.2	631	1	1
Fix_C	2123	627.5	47.3	631	1	1
w/o Fix_C	2123	622.1	56.9	631	4	1
Mathematics						
Target	2196	610.4	54.2	615	3	2
Fix_C	2196	610.6	53.1	616	3	2
w/o Fix_C	2196	609.6	55.7	616	3	2
Grade 7						
Reading						
Target	1923	647.8	52.4	656	4	0
Fix_C	1923	647.7	52.3	655	4	0
w/o Fix_C	1923	646.9	53.1	656	5	0
Mathematics						
Target	1899	650.1	57.0	658	5	0
Fix_C	1899	651.1	52.1	657	4	0
w/o Fix_C	1899	649.9	57.9	658	5	0
Grade 11						
Reading						
Target	2672	682.5	48.0	685	3	0
Fix_C	2672	681.6	47.7	684	2	0
w/o Fix_C	2672	678.2	53.5	685	6	0
Mathematics						
Target	2554	705.3	59.5	711	5	1
Fix_C	2554	704.5	60.3	711	5	1
w/o Fix_C	2554	701.1	65.2	711	9	1

Note. Fix_C = fixing C-parameters for anchor items; w/o Fix_C = without fixing C-parameters for anchor items; LOSS = lowest obtainable scale score; HOSS = highest obtainable scale score.

We found some differences in means and standard deviations of student scale scores in grade 4 Reading Comprehension and grade 11 Reading Comprehension and Mathematics tests. However, the medians of student scale scores for both Fix_C and w/o Fix_C and the target for these grades and contents were very similar. Thus, the mean difference between the w/o Fix_C and the target seems to be caused by the differences in extreme score scales, not in the middle score scale. In addition to this inference, by looking at LOSS and HOSS statistics, the difference of means is likely to be related to the difference in lower score scale. To confirm this inference, the scale scores at several selected percentiles are presented in Table 6.

As would be expected, there were large differences in

Table 6

Scale Scores Corresponding Percentiles for With and Without Fixing C-Parameter Procedures

	5	10	20	30	40	50	60	70	80	90	95
Grade 4											
Reading											
Target	553	570	594	609	620	631	642	653	665	682	699
Fix_C	548	566	592	608	621	631	642	651	663	678	692
w/o Fix_C	526	562	593	609	621	631	641	650	660	675	687
Mathematics											
Target	531	556	581	595	606	615	625	635	646	661	677
Fix_C	529	556	580	595	606	616	625	635	646	661	678
w/o Fix_C	523	556	582	596	606	616	625	635	646	660	677
Grade 7											
Reading											
Target	543	582	613	632	645	656	666	676	686	702	717
Fix_C	539	583	614	631	645	655	667	676	686	702	715
w/o Fix_C	525	581	613	632	645	656	666	676	686	702	714
Mathematics											
Target	527	587	615	632	646	658	669	679	692	710	724
Fix_C	533	589	617	633	646	657	668	679	692	710	723
w/o Fix_C	521	584	616	633	647	658	669	680	692	710	725
Grade 11											
Reading											
Target	606	624	646	662	675	685	696	706	720	739	757
Fix_C	603	622	646	662	674	684	695	706	720	738	754
w/o Fix_C	546	611	644	661	674	685	695	706	720	738	753
Mathematics											
Target	566	632	665	683	699	711	724	736	750	771	786
Fix_C	562	626	663	682	698	711	724	737	751	770	786
w/o Fix_C	560	592	659	681	697	711	724	737	751	771	786

Note. Fix_C = fixing C-parameters for anchor items; w/o Fix_C = without fixing C-parameters for anchor items.

terms of student scale scores in the lower score scale such as 5th and 10th percentiles in the grade 4 Reading Comprehension and grade 11 Reading Comprehension and Mathematics tests. The maximum difference was 60 scale score points at 5th percentile between the w/o Fix_C and target in a grade 11 Reading Comprehension test. However, there was just 3 scale score points difference between the Fix_C and the target in this case. Even though the differences between the w/o Fix_C and the target were evident in these grades and content areas, for other grades and contents the Fix_C produced scale scores that were a little closer to those of the target.

Discussion

This study was designed to investigate the performance of a new approach to fixing C-parameters for anchor items (Fix_C) to be used in the Stocking and Lord equating. This new approach was compared to the conventional approach which does not fix C-parameters for anchor items (w/o Fix_C). Comparisons were made with respect to calibration, equating, and scoring students.

Both the Fix_C and w/o Fix_C approaches had similar characteristics in calibration in terms of the number of EM cycles, model-data fits, non-converging items, and other aspects. Additionally, we obtained very similar estimates of all A, B, and C parameters from two approaches for non-anchor items. However, non-negligible differences between the two approaches were found in estimates for anchor items, especially in C-parameters. This implies that fixing C-parameters for anchor items does not affect the parameter estimation for non-anchor items, but it could have some effects on the SL equating that uses parameter estimates of anchor items.

The Stocking and Lord's F, which can be utilized as a criterion to evaluate the quality of SL equating, indicated that the equating functions of the Fix_C approach were, at least, as good as those of the w/o Fix_C approach. In some cases, the Fix_C produced much better equating functions than did the w/o Fix_C. The comparison of test characteristic curves showed that poor equating of the w/o Fix_C could be related to the non-negligible difference of TCC's in the lower score scale. The analyses of comparing equated item parameters led us to the same conclusion.

The mean and standard deviation of student scale scores from the Fix_C approach were more similar to those of the target than were those of the w/o Fix_C approach. In some grades and in terms of some content, there were large differences between the w/o Fix_C and the target. However, the Fix_C, w/o Fix_C, and the target produced almost the same medians across grades and contents. We can infer from this fact that the mean difference is likely to result from the differences in extreme score scales, not in the middle score scale. By examining scale scores corresponding to several selected percentiles, large differences in scale scores were found in the lower score scale between the w/o Fix_C and the target.

In conclusion, the conventional approach of SL equating, which does not fix C-parameters for anchor items may produce poor equating functions, which are caused by the poor estimation of C-parameters for anchor items in an equating context. The resultant poor equating will lead to biases in item parameter estimates and inaccurate student ability estimates, which in turn, could lead to misinterpretations of test scores. These problems will be more severe in estimating student abilities that are located in the lower score scale. The equating functions of the new approach introduced in this study are, at least, as good as those of the conventional approach, and in some cases they are much better. Therefore, we would recommend their use as a new approach to fix C-parameters for anchor items in IRT equating. The relative appropriateness of IRT equating seems related to the accuracy of C-parameter estimates. Thus, our recommendation would be more logical when C-parameter estimates of anchor items from old and new forms are to some extent, different.

References

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bucket, G. R. (1996). *PARDUXMX* [Computer program]. Unpublished.

Bucket, G. R. (2000). *PARDUXMJ* [Computer program]. Unpublished.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practices, 10*, 37-45.

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.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hanson, B. A., & Beguin, 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.

Hung, P., Wu, Y., & Chen, Y. (1991). *IRT item parameter linking: Relevant issues for the purpose of item banking*. Paper presented at the International Academic Symposium on Psychological Measurement, Tainan, Taiwan.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51-66.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York : Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

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.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement, 25,* 3-24.

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

Thissen, D. M., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397-412.

Way, W. D., & Tang, K. L. (1991). *A comparison of four logistic model equating methods.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST users guide*. Princeton, NJ: Educational Testing Service.

Yen, W. (1990). *CTB scaling specifications*. Unpublished research paper.

Received October 16, 2006
Revision received December 13, 2007
Accepted January 19, 2008