DOCUMENT RESUME

ED 333 050 TM 016 648

AUTHOR Kim, Seock-Ho; Cohen, Allan S.
TITLE Effects of Linking Methods on Detection of DIF.
PUB DATE Apr 91
NOTE 33p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Chi Square; Comparative Analysis; Equations (Mathematics); Estimation (Mathematics); *Evaluation Methods; *Item Bias; Item Response Theory; *Mathematical Models; *Sample Size; Simulation
IDENTIFIERS Item Characteristic Function; Item Parameters; *Linking Metrics

ABSTRACT

Studies of differential item functioning (DIF) under item response theory require that item parameter estimates be placed on the same metric before comparisons can be made. Evidence that methods for linking metrics may be influenced by the presence of differentially functioning items has been inconsistent. The effects of three methods for linking metrics on detection of DIF were studied. The methods included: (1) a weighted mean and sigma method; (2) the test characteristic curve method; and (3) the minimum chi-square method. Both iterative and non-iterative linking procedures were compared for each method. A two-parameter logistic item characteristic model was used to generate eight simulated data sets. A 60-item test and sample sizes of 300 and 600 were generated for each data set. Results indicate that detection of DIF following linking via the test characteristic curve method gave the most accurate results when the sample size was small. When the sample size was large, results for the three methods were essentially the same. Iterative linking provided a substantial improvement in detection of DIF over non-iterative linking. A 28-item list of references is included. Seven tables present study findings, and an appendix summarizes the three methods. (Author/SLD)

Effects of Linking Methods on Detection of DIF

Seock-Ho Kim and Allan S. Cohen

University of Wisconsin-Madison

Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL

April, 1991

Running Head: EFFECTS OF LINKING METHODS

Abstract

Studies of differential item functioning under item response theory require that item parameter estimates be placed on the same metric before comparisons can be made. Evidence that methods for linking metrics may be influenced by the presence of differentially functioning items has shown inconsistent findings. The present study compared the effects of three methods for linking metrics, a weighted mean and sigma method, the test characteristic curve method, and the minimum chi-square method, on detection of differential item functioning. Both iterative and noniterative linking procedures were compared for each method. Results indicated that detection of differentially functioning items following linking via the test characteristic curve method gave the most accurate results when the sample size was small. When the sample size was large, results for the three linking methods were essentially the same. Iterative linking provided a substantial improvement in detection of differentially functioning items over noniterative linking.

Index terms: differential item functioning, equating, item response theory, iterative linking, linking metrics.

Effects of Linking Methods on Detection of DIF

Studies of differential item functioning (DIF) under item response theory (IRT) require that item parameter estimates obtained on the same item but from different samples of examinees are expressed in the same metric. Current item parameter estimation methods yield metrics which are unique only up to a linear transformation. To link both sets of estimates, that is, to place them both on the same metric, it is necessary to determine the slope and intercept of the linear equation required for the transformation. The present investigation was designed to examine the effects of linking method on detection of DIF under IRT.

Two general classes of linking methods have been developed for this purpose: mean and sigma methods and characteristic curve methods. Mean and sigma methods use the first two moments of the distribution of item difficulty estimates to determine the appropriate linear equation (cf., Bejar & Wingersky, 1981; Cook, Eignor, & Hutten, 1979; Linn, Levine, Hasting, & Wardrop, 1981; Loyd & Hoover, 1980; Marco, 1977; Vale, 1986). One problem with these methods is that they do not use information available from the estimated item discrimination parameters in obtaining the linking equation. In contrast, characteristic curve methods (cf., Divgi, 1980; Haebara, 1980; Stocking & Lord, 1983) do make use of the information available from both the item discrimination and item difficulty parameters. This second class of methods derives a linking equation by minimizing some measure of the

difference between the test characteristic curves estimated in each sample. The minimum chi-square method (Divgi, 1985) is a variation of the characteristic curve method in which the standard errors of the estimates of the item parameters are included in the linear equation.

Comparisons between the two types of linking methods have not been conclusive. Stocking and Lord (1983) reported that the characteristic curve method was more accurate than the robust iterative weighted mean and sigma method similar to that reported by Linn et al. (1981). Baker and Al-Karni (in press), however, noted no differences between the unweighted mean and sigma method of Loyd and Hoover (1980) and the Stocking and Lord characteristic curve procedure. Candell and Drasgow (1988) found that the weighted mean and sigma procedure of Linn et al. (1981) was more accurate than the characteristic curve method. No studies have yet compared the accuracy of the minimum chi-square method with these other methods. Since the minimum chi-square method combines the information used in the mean and sigma method with that used in the characteristic curve method, one assumption is that this method will be more accurate than either of the other two methods alone.

Linking methods may be seriously affected by the presence of DIF items such that errors in the linking transformation may result in spurious identification of items as functioning differentially (Shepard, Camilli, & Williams, 1984). Lord (1980) outlined the following procedure, suggested by Marco,

for reducing the potential effects of DIF items on parameter estimation and subsequent detection of DIF:

1. Estimate item parameters for all groups combined, standardizing on item difficulty estimates.

2. Re-estimate item parameters for each group separately holding the guessing parameters, standardizing on item difficulty estimates.

3. Identify DIF items and remove them.

4. Combine groups and estimate ability for each examinee.

5. Hold ability fixed and re-estimate item difficulty and discrimination for all items for each group separately.

6. Identify DIF items.

One problem with this approach is that it requires re-estimation of item and ability parameters. Candell and Drasgow (1988) reported the following alternative procedure, due to Segall (1983), which is somewhat easier to implement:

1. Estimate item parameters independently in each group.

2. Link metrics across groups.

3. Estimate DIF indices and remove DIF items.

4. Relink group metrics using only non-DIF items.

5. Re-estimate DIF indices and remove DIF items.

Steps 4 and 5 are continued until either no DIF items are detected or the same items are identified as DIF items on two successive iterations. The Candell and Drasgow procedure is somewhat quicker and easier to implement in that parameters do not have to be re-estimated. This approach to iterative linking was used in the present study.

The final linking is based only on items which are not identified as DIF items. Following this linking, DIF indices are then re-estimated for all items. Clearly, iterative linking is costly, although results indicate that DIF detection is improved over noniterative linking (Candell & Drasgow, 1988; Candell & Hulin, 1986; Drasgow, 1987; Hulin & Mayer, 1986; Kok, Mellenbergh, & van der Flier, 1985; van der Flier, Mellenbergh, Adèr, & Wijn, 1984).

The present study was designed to compare the effectiveness of three linking methods on the detection of DIF: (1) the weighted mean and sigma method (WMS) (Linn et al., 1981); (2) the test characteristic curve method (TCC) (Stocking & Lord, 1983); and (3) the minimum chi-square method (MCS) (Divgi, 1985). In addition, the relative contributions of iterative and noniterative linking to the detection of DIF for different sample sizes were also examined.

Methods

Data Generation

A two-parameter logistic item characteristic model (2PM) was used to generate eight simulated data sets using the computer program GENIRV (Baker, 1978). In this model, the probability of an examinee i giving a correct response for item j is a function of the discrimination of the item, a_j, the difficulty of the item, b_j, and the examinee's unidimensional ability, θ_i. This probability is expressed as

$$P_j(\theta_i) = P(\theta_i, a_j, b_j) = [1 + \exp\{-Da_j(\theta_i - b_j)\}]^{-1}, \qquad (1)$$

where D is a scaling constant equal to 1 or 1.702. In this study, item and ability parameters were expressed in the logistic metric (i.e., $D = 1$).

A 60-item test was generated in each data set. Generating parameters for the underlying ability distributions for both the reference and focal groups were normal (0, 1). The generated item discrimination parameters were lognormally distributed (0, .25); that is, log a_j was distributed as normal (0, .25). The generated item difficulty parameters were matched to the θ distribution and distributed normal (0, 1). Item parameters used to generate the data sets are given in Table 1.

****

Insert Table 1 About Here

****

Hullin, Lissak, and Drasgow (1982) recommend a minimum of 500 examinees for the 2PM. Data sets were generated using two sample sizes, 300 and 600 examinees, to permit a comparison of the effects of sample size. Candell and Drasgow (1988) suggest that the number of DIF items may affect the metric used for linking. For each sample size, therefore, one reference group (R) and three focal groups (F) were generated. Three proportions of DIF items were used in the simulated tests for the Focal groups: 0 percent (Focal-0, F0); 10 percent (Focal-10, F10); and 20 percent (Focal-20, F20). Two types of DIF were simulated in the present study: uniform DIF (for which $a_R = a_F$ and $b_R \neq b_F$) and non-uniform DIF (for which $a_R \neq a_F$ and either $b_R \neq b_F$ or $b_R \neq b_F$).

Parameter Estimation

Item and ability parameters were estimated via the computer program BILOG 3 (Mislevy & Bock, 1990). BILOG 3 default conditions implement a marginal Bayesian estimation (i.e., marginal maximum a posteriori estimation) procedure for the 2PM. Previous research (Mislevy & Bock, 1986; Mislevy & Stocking, 1989) has suggested that, when sample sizes are small, the marginal Bayesian estimation approach appears to provide estimates which are closer to the underlying values of the generating distributions than those obtained via other estimation procedures.

Linking Methods

Under the assumption of item parameter invariance, item parameters

estimated in different groups will differ by a linear transformation from one group to another. Thus, it is possible to equate the metrics from these different groups by means of some linear transformation so that between groups comparisons of parameters can be made.

In the DIF study context, estimates from the calibration of the focal group are transformed to the metric of the reference group. The transformed estimates of item discrimination and difficulty parameters for item j are given by

$$a^*_{jF} = a_{jF}/A \tag{2}$$

and

$$b^*_{jF} = Ab_{jF} + B, \tag{3}$$

where * indicates a transformed value. Further, the estimate of the ability parameter for examinee i in the focal group can be converted to the reference group scale using

$$\theta^*_{iF} = A\theta_{iF} + B. \tag{4}$$

The task is to determine appropriate coefficients A and B. A brief description of the three linking methods used in this study for determining these equating constants is presented in the Appendix.

Iterative Linking

The iterative linking procedure described by Candell and Drasgow (1988) was used for the present study. Results for the noniterative conditions were obtained from the first iteration of each linking method.

Measurement of DIF

A χ^2 statistic developed by Lord (1980) was used to detect DIF. Lord's χ^2 simultaneously tests the hypothesis that a_j and b_j are identical across groups. This is the same statistic as Q_j which appears in the MCS method. Evidence has been presented (McLaughlin & Drasgow, 1987) which indicates that the Type I error rate of this statistic may be seriously violated under certain conditions. Since the data for this study were simulated, the Type I error rates could be evaluated.

Results

Recovery of Item and Ability Parameters

A recovery analysis was done to determine the extent to which the generating parameters were captured in the simulated data sets. The root mean squared differences (RMSD) and correlations (r) between generating parameters and parameter estimates are given in Table 2 for each of the eight data sets. To assess the adequacy of parameter recovery, the calibrations must all be transformed to a common metric (Baker & Al-Karni, in press; Yen, 1987). In the present study, the parameter estimates were transformed to the underlying metric using the TCC method.

****

Insert Table 2 About Here

****

The RMSDs for the equated values of item difficulty and discrimination in the small sample were larger than those in the larger sample. Recovery, as represented by the RMSDs, was better in the larger sample (i.e., 600 examinees). The same tendency appeared in the correlations between the item discrimination estimates and the underlying values. The correlations for item difficulty and ability showed no differences across all eight data sets. Based on these results, recapture of the underlying item and ability parameters appeared to be very good.

Detection of DIF

Results of the DIF detection procedures are given in Tables 3 and 4 for the small and large sample conditions, respectively. The effectiveness of DIF detection can be seen clearly in the number of false positive (FP) and false negative (FN) items identified under each linking method. Results for the small sample condition (i.e., 300 examinees) and large sample condition were similar. No FN items were observed under the null condition (i.e., RF0) for either alpha level. FN identifications (i.e., failures to detect the DIF items) did occur when DIF items were present on the test. Furthermore, the number of FN identifications increased with an increase in the percentage of DIF items on the test. The number of FP identifications (i.e., incorrectly identifying an item as a DIF item) did not appear to be related to the percentage of DIF items on the test, although it clearly increased as the alpha level increased from .01 to .05. Sample size differences were apparent in that the

number of FN identifications was lower for the larger sample.

****

Insert Tables 3 and 4 About Here

****

The results from the final iteration, given in Tables 5 and 6 for each linking method, indicate FN identifications occurred with the same frequency for all linking method under each condition for each sample size. This was generally true for the FP identifications as well. More FP items tended to be identified when the WMS method was used in the presence of DIF. The fewest FP identifications were generally made following TCC linking, particularly in the small samples.

****

Insert Tables 5 and 6 About Here

****

The presence of FN identifications is always a major problem in any DIF detection study as these are the items which pose a major threat to validity. It is of interest to note that, when FN identifications did occur, all three linking methods had the same results. Examination of these items revealed that the majority were nonuniform DIF items in which the generating item difficulties were equal in both the reference and focal groups (i.e., $a_R \neq a_F, b_R = b_F$). In fact, of the four items containing this type of DIF (items 5, 10, 25 and 30) under RF20, none were detected in the small samples for the .01 nominal

alpha level and only one item (30) was detected for the .05 alpha level. In the large samples, two items (5 and 10) were not detected as DIF items at both alpha levels. The other two FN items for the small samples (55, 60) were also nonuniform DIF items but had $b_R \neq b_F$. These items were likely missed as the values of the item difficulties were 2 10. That is, they were far from the center of the ability distribution and, consequently, had larger standard errors.

Iterative vs. Noniterative Linking

Iterative linking had a consistent effect on the estimation of the A constant for the WMS method. The change from first iteration to the last occurred in estimates of A from both large and small samples. Under the null condition, the change was a very slight decrease. When DIF items were present, however, the estimate of A increased with iterations. No such change was observed for the other two linking methods. The number of iterations was lower for the larger sample sizes for all linking methods.

The effect of the use of an iterative procedure also can be seen by comparison of the numbers of FP and FN identifications on the first and last iterations. These results were similar for both the small and large sample conditions. The FN rate changed only very slightly from first iteration to last across all linking methods. The FP rate, however, did decrease with the use of iterative linking procedures. This decrease occurred primarily for the WMS method. A similar decrease occurred for the TCC and MCS methods

only under the RF20 condition for $\alpha = .05$.

Comparison of Linking Methods

The choice of linking method did not appear to have an effect on the rate of FN identifications. There were some notable differences, however, in the numbers of FP identifications among the linking methods. The FP rates were nearly always higher for the WMS method. This was particularly evident in the small sample results.

****

Insert Table 7 About Here

****

Correlations between values of the DIF statistic, Lord's χ^2, following the final iteration provide another indication of the degree of similarity in the results from each linking method. The correlations in Table 7 are all high indicating substantial similarity among linking methods. There were no real differences in correlations for the large sample; all were essentially perfect. For the small sample, correlations for the RF10 and RF20 conditions were also quite high, although under the RF0 condition, the correlations for the TCC and MCS methods were relatively higher.

Discussion

The presence of DIF in a test item is a serious problem affecting the validity of that item as well as of the entire test. If procedures developed to

detect DIF are themselves influenced by the particular linking method used, then the detection of DIF is also likely to suffer. The results of the present study give some indication of the differences in detection of DIF associated with the particular method used to link metrics.

For small samples, the TCC method generally provided the most accurate detection of DIF, particularly when iterative linking was used. Detection under MCS linking was nearly as accurate. The TCC method also provided more accurate linking for both iterative and noniterative linking when no DIF items were present. This result is in disagreement with the findings of Candell and Drasgow (1988) who reported that the WMS method provided more accurate results than the TCC method. The MCS method performed about as effectively under these conditions as the TCC method.

There were no real differences in DIF detection related to linking methods in the large sample conditions. In fact, the detection of DIF items under both TCC and MCS linking transformations was not substantially different for the different sample sizes. This is somewhat surprising as one would expect standard errors of item parameter estimates to decrease with an increase in sample size. This reduction would, in turn, yield an improvement in the accuracy of the transformation. If DIF detection is related to the accuracy of the linking transformation, therefore, one would expect a subsequent improvement in detection of DIF when standard errors of item parameter estimates are decreased. As the MCS method combines the information used

by both the WMS and TCC methods, one would expect the MCS method to yield better detection of DIF. In fact, this did not occur.

The presence of DIF items clearly tended to increase the number of FP items for the first iteration. This was particularly evident with the WMS method. McCauley and Mendoza (1985) and Candell and Drasgow (1988) reported similar results. In each case, iterative linking resulted in a decrease in the number of FP and, to a lesser extent, FN identifications.

Choice of a linking method appears to be important primarily in the context of small sample sizes. This is often the case, for example, with DIF detection studies in which a focal group is a minority group. From an implementation point of view, the WMS method is the easiest to adopt as programming of this method is relatively simple and straight forward. The MCS method is also relatively simple to implement. The TCC method, however, is more difficult as it requires development of some difficult programming code. If software is available for each of the methods, the results of the present study would mitigate in favor of selection of the TCC method. If sample sizes are large, choice of one of these three methods does not seem as critical.

References

Baker, F. B. (1978). *GENIRV: A program to generate item response vectors*. Unpublished manuscript, University of Wisconsin–Madison, Laboratory of Experimental Design.

Baker, F. B. (1990). *EQUATE: Computer program for equating two metrics in item response theory*. University of Wisconsin–Madison, Laboratory of Experimental Design.

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

Bejar, I. I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English*. College Board Report No. 81-8. Princeton, NJ: Educational Testing Service.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253-260.

Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons: Independent sources of information about item non-equivalence. *Journal of Cross-Cultural Psychology*, *17*, 417-440.

Cook, L. L., Eignor, D. R., & Hutten, L. R. (1979, April). *Considerations in the application of latent trait theory to objectives-based criterion-referenced tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Divgi, D. R. (1980, April). *Evaluation of scales for multilevel test batteries*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement*, *9*, 413-415.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19-29.

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

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*, 249-260.

Hulin, C. L., & Mayer, L. (1986). Psychometric equivalence of a translation of the job descriptive index into Hebrew. *Journal of Applied Psychology*, *71*, 83-94.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement*, *22*, 295-303.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159-173.

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

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.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement*, *11*, 161-173.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* [computer program]. Mooresville,

IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57-75.

Segall, D. O. (1983). *Test characteristic curves, item bias, and transformation to a common metric in item response theory: A methodological artifact with serious consequences and a simple solution* Unpublished manuscript, University of Illinois, Department of Psychology.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, *9*, 93-128.

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

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*, 333-344.

van der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement*, *21*, 131-145.

Yen, W. M. (1987). A comparison of the efficiency and accuracy to BILOG and LOGIST. *Psychometrika*, *52*, 275-291.

Table 1
Item Parameters Used to Generate the Data Sets

	Data Set					
	Reference[a]		Focal-10[b]		Focal-20[c]	
Item No.	Disc.	Diff.	Disc.	Diff.	Disc.	Diff.
1	0.81	-2.10	0.81	-2.10	0.81	-2.10
2	1.00	-2.10	1.00	-2.10	1.00	-2.10
3	1.23	-2.10	1.23	-2.10	1.23	-2.10
4	0.66	-1.40	0.66	-1.40	0.66	-1.40
5	0.81	-1.40	0.81	-1.40	(0.65)[d]	-1.40
6	1.00	-1.40	1.00	-1.40	1.00	-1.40
7	1.00	-1.40	1.00	-1.40	1.00	-1.40
8	1.23	-1.40	1.23	-1.40	1.23	-1.40
9	1.52	-1.40	1.52	-1.40	1.52	-1.40
10	0.53	-0.70	(0.37)	-0.70	(0.37)	-0.70
11	0.66	-0.70	0.66	-0.70	0.66	-0.70
12	0.81	-0.70	0.81	-0.70	0.81	-0.70
13	0.81	-0.70	0.81	-0.70	0.81	-0.70
14	1.00	-0.70	1.00	-0.70	1.00	-0.70
15	1.00	-0.70	1.00	-0.70	1.00	(-0.20)
16	1.00	-0.70	1.00	-0.70	1.00	-0.70
17	1.00	-0.70	1.00	-0.70	1.00	-0.70
18	1.23	-0.70	1.23	-0.70	1.23	-0.70
19	1.23	-0.70	1.23	-0.70	1.23	-0.70
20	1.52	-0.70	1.52	(-0.20)	1.52	(-0.20)
21	1.88	-0.70	1.88	-0.70	1.88	-0.70
22	0.53	0.00	0.53	0.00	0.53	0.00
23	0.66	0.00	0.66	0.00	0.66	0.00
24	0.66	0.00	0.66	0.00	0.66	0.00
25	0.81	0.00	0.81	0.00	(0.49)	0.00
26	0.81	0.00	0.81	0.00	0.81	0.00
27	0.81	0.00	0.81	0.00	0.81	0.00
28	0.81	0.00	0.81	0.00	0.81	0.00
29	1.00	0.00	1.00	0.00	1.00	0.00
30	1.00	0.00	(0.68)	0.00	(0.68)	0.00

Table 1-continued
Item Parameters Used to Generate the Data Sets

	Data Set					
	Reference[a]		Focal-10[b]		Focal-20[c]	
Item No.	Disc.	Diff.	Disc.	Diff.	Disc.	Diff.
31	1.00	0.00	1.00	0.00	1.00	0.00
32	1.00	0.00	1.00	0.00	1.00	0.00
33	1.23	0.00	1.23	0.00	1.23	0.00
34	1.23	0.00	1.23	0.00	1.23	0.00
35	1.23	0.00	1.23	0.00	1.23	(1.00)
36	1.23	0.00	1.23	0.00	1.23	0.00
37	1.52	0.00	1.52	0.00	1.52	0.00
38	1.52	0.00	1.52	0.00	1.52	0.00
39	1.88	0.00	1.88	0.00	1.88	0.00
40	0.53	0.70	0.53	(1.70)	0.53	(1.70)
41	0.66	0.70	0.66	0.70	0.66	0.70
42	0.81	0.70	0.81	0.70	0.81	0.70
43	0.81	0.70	0.81	0.70	0.81	0.70
44	1.00	0.70	1.00	0.70	1.00	0.70
45	1.00	0.70	1.00	0.70	(0.68)	(1.70)
46	1.00	0.70	1.00	0.70	1.00	0.70
47	1.00	0.70	1.00	0.70	1.00	0.70
48	1.23	0.70	1.23	0.70	1.23	0.70
49	1.23	0.70	1.23	0.70	1.23	0.70
50	1.52	0.70	(1.20)	(1.70)	(1.20)	(1.70)
51	1.88	0.70	1.88	0.70	1.88	0.70
52	0.66	1.40	0.66	1.40	0.66	1.40
53	0.81	1.40	0.81	1.40	0.81	1.40
54	1.00	1.40	1.00	1.40	1.00	1.40
55	1.00	1.40	1.00	1.40	(0.68)	(2.40)
56	1.23	1.40	1.23	1.40	1.23	1.40
57	1.52	1.40	1.52	1.40	1.52	1.40
58	0.81	2.10	0.81	2.10	0.81	2.10
59	1.00	2.10	1.00	2.10	1.00	2.10
60	1.23	2.10	(0.91)	(3.10)	(0.91)	(3.10)

[a] Focal-0 has the same item parameters.
[b] Focal-10 contains two uniform DIF items (20, 40) and four non-uniform DIF items (10, 30, 50, 60).
[c] Focal-20 contains four uniform DIF items (15, 20, 35, 40) and eight non-uniform DIF items (5, 10, 25, 30, 45, 50, 55, 60).
[d] ( ) indicates values different from reference group.

Table 2
Root Mean Squared Differences (RMSD) and Correlation (r)
Between Estimates and True Values

Examinee	Group	Discrimination		Difficulty		Ability	
		RMSD	(r)	RMSD	(r)	RMSD	(r)
300	Reference	.1517	(.9437)	.1183	(.9943)	.2098	(.9784)
	Focal-0	.1761	(.8775)	.1414	(.9903)	.2049	(.9795)
	Focal-10	.1673	(.8754)	.1449	(.9910)	.2121	(.9783)
	Focal-20	.1549	(.9019)	.1517	(.9909)	.2236	(.9756)
600	Reference	.1049	(.9595)	.1011	(.9950)	.2109	(.9775)
	Focal-0	.0875	(.9708)	.0915	(.9959)	.2032	(.9792)
	Focal-10	.1190	(.9399)	.1067	(.9952)	.2067	(.9785)
	Focal-20	.1066	(.9524)	.1166	(.9950)	.2170	(.9765)

Note. Estimates were transformed to the true metric using TCC method.

Table 3
Equating Constants and Number of False Positive (FP) and False Negative (FN) Items on Each Iteration for 300 Examinees

	1st Iteration		2nd Iteration		3rd Iteration		4th Iteration	
Method[a]	Constants	FP—FN	Constants	FP—FN	Constants	FP—FN	Constants	FP—FN
RF0-.01-WMS	A=0.9731	1—0	A=0.9598	1—0				
	B=-.0007		B=-.0008					
TCC	A=1.0254	0—0						
	B=-.0293							
MCS	A=0.9949	1—0	A=0.9961	1—0				
	B=-.0397		B=-.0393					
.05-WMS	A=0.9731	2—0	A=0.9571	4—0	A=0.9465	4—0		
	B=-.0007		B=-.0008		B=-.0006			
TCC	A=1.0254	2—0	A=1.0226	2—0				
	B=-.0293		B=-.0377					
MCS	A=0.9949	3—0	A=0.9965	4—0	A=0.9927	4—0		
	B=-.0397		B=-.0335		B=-.0269			
RF10-.01-WMS	A=0.8985	2—3	A=0.9994	1—3	A=1.0052	1—3		
	B=-.0014		B=-.0009		B=-.0009			
TCC	A=0.9891	1—3	A=0.9971	1—3				
	B=-.0910		B=-.0654					
MCS	A=0.9940	1—3	A=0.9885	1—3				
	B=-.0741		B=-.0521					
.05-WMS	A=0.8985	8—2	A=0.9570	5—2	A=0.9898	4—1	A=1.0003	4—1
	B=-.0014		B=0.0002		B=-.0005		B=-.0007	
TCC	A=0.9891	1—1	A=1.0080	1—1				
	B=-.0910		B=-.0610					
MCS	A=0.9940	1—1	A=0.9954	1—1				
	B=-.0741		B=-.0520					
RF20-.01-WMS	A=0.8335	3—6	A=0.9788	0—6	A=1.0151	0—6		
	B=-.0023		B=-.0004		B=-.0008			
TCC	A=0.9646	0—6	A=0.9913	0—6				
	B=-.1202		B=-.0481					
MCS	A=0.9702	0—6	A=0.9876	0—6				
	B=-.1012		B=-.0351					
.05-WMS	A=0.8335	5—6	A=0.9678	2—5	A=0.9640	2—5		
	B=-.0023		B=-.0004		B=-.0019			
TCC	A=0.9646	5—5	A=0.9930	0—5	A=0.9999	1—5	A=0.9958	1—5
	B=-.1202		B=-.0718		B=-.0504		B=-.0449	
MCS	A=0.9702	2—5	A=0.9882	1—5	A=0.9904	1—5		
	B=-.1012		B=-.0441		B=-.0313			

[a] Data Set-Alpha-Method

Table 4
Equating Constants and Number of False Positive (FP) and False Negative (FN) Items on Each Iteration for 600 Examinees

	1st Iteration		2nd Iteration		3rd Iteration		4th Iteration	
Method*	Constants	FP—FN	Constants	FP—FN	Constants	FP—FN	Constants	FP—FN
RF0-.01-WMS	A=1.0077	0—0						
	B=-.0003							
TCC	A=1.0080	0—0						
	B=-.0086							
MCS	A=1.0081	0—0						
	B=-.0065							
.05-WMS	A=1.0077	4—0	A=0.9992	4—0				
	B=-.0003		B=-.0005					
TCC	A=1.0080	4—0	A=1.0040	4—0				
	B=-.0086		B=-.0271					
MCS	A=1.0081	4—0	A=1.0020	4—0				
	B=-.0065		B=-.0221					
RF10-.01-WMS	A=0.9132	0—3	A=1.0265	0--0	A=1.0321	0—0		
	B=-.0009		B=-.0001		B=-.0001			
TCC	A=0.9794	0—2	A=1.0032	0—0	A=1.0193	0—0		
	B=-.0433		B=-.0046		B=0.0032			
MCS	A=1.0023	0—0						
	B=-.0321							
.05-WMS	A=0.9132	5—0	A=1.0147	1—0	A=1.0309	1—0		
	B=-.0009		B=-.0001		B=0.0001			
TCC	A=0 9794	2—0	A=1.0141	1—0	A=1.0188	1—0		
	B=-.0433		B=0.0001		B=0.0081			
MCS	A=1.0023	2—0	A=1.0147	1—0	A=1.0198	1—0		
	B=-.0321		B=-.0007		B=0.0076			
RF20-.01-WMS	A=0.8886	3—5	A=1.0091	2—4	A=1.0145	2—4		
	B=-.0020		B=-.0003		B=-.0001			
TCC	A=0.9555	2—4	A=0.9861	2—4	A=0.9864	2—4		
	B=-.0901		B=-.0175		B=-.0060			
MCS	A=0.9959	1—4	A=1.0082	1—4				
	B=-.0801		B=-.0102					
.05-WMS	A=0.8886	7—4	A=1.0090	4--2	A=1.0299	4—2		
	B=-.0020		B=-.0003		B=-.0001			
TCC	A=0.9555	9—3	A=0.9811	5—2	A=1.0005	5—2		
	B=-.0901		B=-.0441		B=-.0119			
MCS	A=0.9959	7—3	A=1.0106	4—2	A=1.0154	4—2		
	B=-.0801		B=-.0193		B=-.0102			

* Data Set-Alpha-Method

Table 5
Number of False Positive (FP) and False Negative (FN) Items and Their Locations on Final Iteration for 300 Examinees

Data Set	Alpha	Method	Iteration	FP	(Item No.)	FN	(Item No.)
RF0	.01	WMS	2	1	(55)	0	
		TCC	1	0		0	
		MCS	2	1	(55)	0	
	.05	WMS	3	4	(2,9,38,55)	0	
		TCC	2	2	(23,55)	0	
		MCS	3	4	(2,23,38,55)	0	
RF10	.01	WMS	3	1	(55)	3	(10,30,60)
		TCC	2	1	(55)	3	(10,30,60)
		MCS	2	1	(55)	3	(10,30,60)
	.05	WMS	4	4	(2,31,38,55)	1	(60)
		TCC	2	1	(55)	1	(60)
		MCS	2	1	(55)	1	(60)
RF20	.01	WMS	3	0		6	(5,10,25,30,55,60)
		TCC	2	0		6	(5,10,25,30,55,60)
		MCS	2	0		6	(5,10,25,30,55,60)
	.05	WMS	3	2	(9,31)	5	(5,10,25,55,60)
		TCC	4	1	(31)	5	(5,10,25,55,60)
		MCS	2	1	(31)	5	(5,10,25,55,60)

Table 6
Number of False Positive (FP) and False Negative (FN) Items and Their Locations on Final Iteration for 600 Examinees

Data Set	Alpha	Method	Iteration	FP	(Item No.)	FN	(Item No.)
RF0	.01	WMS	1	0		0	
		TCC	1	0		0	
		MCS	1	0		0	
	.05	WMS	2	4	(4,31,47,54)	0	
		TCC	2	4	(4,31,47,54)	0	
		MCS	2	4	(4,31,47,54)	0	
RF10	.01	WMS	3	0		0	
		TCC	3	0		0	
		MCS	1	0		0	
	.05	WMS	3	1	(29)	0	
		TCC	3	1	(29)	0	
		MCS	3	1	(29)	0	
RF20	.01	WMS	3	2	(32,54)	4	(5,10,25,60)
		TCC	3	2	(32,54)	4	(5,10,25,60)
		MCS	2	1	(54)	4	(5,10,25,60)
	.05	WMS	3	4	(4,22,32,54)	2	(5,10)
		TCC	3	5	(4,22,32,41,54)	2	(5,10)
		MCS	3	4	(4,22,32,54)	2	(5,10)

Table 7
Correlations of DIF (Lord's Chi-Square) Measures Among Linking Methods Following Final Iteration

		300 Examinees			600 Examinees		
Data Set	Alpha	Method	WMS	TCC	Method	WMS	TCC
RF-0	.01	TCC	.908		TCC	.996	
		MCS	.942	.990	MCS	.998	1.000
	.05	TCC	.873		TCC	.962	
		MCS	.945	.982	MCS	.974	.999
RF10	.01	TCC	.984		TCC	1.000	
		MCS	.990	.999	MCS	.998	.998
	.05	TCC	.986		TCC	1.000	
		MCS	.990	.999	MCS	1.000	1.000
RF20	.01	TCC	.998		TCC	1.000	
		MCS	.999	1.000	MCS	1.000	1.000
	.05	TCC	.998		TCC	1.000	
		MCS	.999	1.000	MCS	1.000	1.000

Appendix

Description of Linking Methods Used in This Study

The following is a description of the three linking methods used in the present study:

Weighted Mean and Sigma Method (WMS). The two equating constants are estimated from the first two moments of the distributions of the weighted estimates of item difficulties. The jth weight is the inverse of the larger of the estimated variances (i.e., squared standard errors) of the item difficulty computed from the focal group and the item difficulty computed from the reference group. In this way, items for which the difficulty parameter was poorly estimated for either of the groups are given relatively less weight in determining the equating constants. Specifically, if b_{jF}^{w*} is the weighted item difficulty of item j in the focal group after equating and b_{jF}^{w} is the corresponding value prior to equating, then

$$b_{jF}^{w*} = Ab_{jF}^{w} + B, \tag{5}$$

where A and B are selected such that the mean and standard deviation of the weighted item difficulties in the focal group are equal to the mean and standard deviation of the weighted item difficulties in the reference group.

For this transformation

$$A = \sigma_{b_R^w}/\sigma_{b_F^w} \tag{6}$$

and

$$B = \mu_{b_R^w} - A\mu_{b_F^w}, \tag{7}$$

where $\mu_{b_R^w}$ is the mean and $\sigma_{b_R^w}$ is the standard deviation of the weighted item difficulties from the reference group and $\mu_{b_R^w}$ and $\sigma_{b_F^w}$ are the corresponding values from the focal group.

Test Characteristic Curve Method (TCC). The TCC method is based on matching the test characteristic curves yielded by calibrations in the reference and focal groups. Let T_{iF} be the true score on the reference group scale for examinee i from the focal group and let T_{iF}^* be the transformed true score for this examinee. Then

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_{iF} - T_{iF}^*)^2 \tag{8}$$

is the quadratic loss-function to be minimized, where N is the number of examinees taking the test. Under the 2PM, T_{iF} and T_{iF}^* are defined as

$$T_{iF} = \sum_{j=1}^{n} P(\theta_{iF}, a_{jR}, b_{jR}) \tag{9}$$

and

$$T_{iF}^* = \sum_{j=1}^{n} P(\theta_{iF}, a_{jF}^*, b_{jF}^*), \tag{10}$$

where n is the number of items used.

The function to be minimized becomes

$$F = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{j=1}^{n} P(\theta_{iF}, a_{jR}, b_{jR}) - \sum_{j=1}^{n} P(\theta_{iF}, a_{jF}^*, b_{jF}^*)\right\}^2. \tag{11}$$

The task, in other words, is to find the values of A and B used to transform T^*_{iF} into T_{iF} that minimizes F.

The mathematics result in two equations in two unknown. Unfortunately, these two equations do not have a closed-form solution. To solve these non-linear equations, Stocking and Lord (1983) employed the multivariate search technique to find the two equating constants that minimize the quadratic loss-function F. In this study, the computer program EQUATE (Baker, 1990), which implements the Stocking and Lord procedure on the IBM-PC, was used to compute equating constants.

Minimum Chi-Square Method (MCS). This method (Divgi, 1985) combines information used in the TCC method with the 2×2 variance-covariance matrix of sampling errors for each item from the item parameter estimation procedure. For item j, let Σ_{jR} and Σ_{jF} be the values of the variance-covariance matrix from the calibrations of the reference and focal groups, respectively. When a_{jF} and b_{jF} are transformed to a^*_{jF} and b^*_{jF}, respectively, the matrix Σ_{jF} is also converted to Σ^*_{jF}, where the diagonal element of Σ_{jF} for the item discrimination (i.e., the squared standard error of the item discrimination) is divided by A^2 and the diagonal element of Σ_{jF} for item difficulty (i.e., the squared standard error of item difficulty) is multiplied by A^2. The quadratic loss-function, Q_j, is calculated as follows:

$$Q_j = \left(a_{jR} - a^*_{jF}, b_{jR} - b^*_{jF}\right) \left(\Sigma_{jR} + \Sigma^*_{jF}\right)^{-1} \left(a_{jR} - a^*_{jF}, b_{jR} - b^*_{jF}\right)' . \quad (12)$$

Let

$$Q = \sum_{j=1}^{n} Q_j \tag{13}$$

and be treated as a function of two equating constants A and B. The task is to find those values of A and B that minimize Q.

Since the partial derivative $\partial Q/\partial B = 0$ is linear with regard to B and easily solved as a function of A, the MCS method can be easily implemented in a computer program than the TCC method which requires the multivariate search technique. Denote $S_{j_{aa}}$ and $S_{j_{bb}}$ are individual elements from the matrix $\mathbf{S}_j = \left(\Sigma_{jR} + \Sigma^{*}_{jF}\right)^{-1}$. Then

$$B = \sum_{j=1}^{n}\{S_{j_{ab}}(a_{jR} - a_{jF}/A) + S_{j_{bb}}(b_{jR} - Ab_{jF})\} / \sum_{j=1}^{n} S_{j_{bb}}. \tag{14}$$

When this value of B is substituted in the expression for Q, we have left a minimization problem with only a single unknown, A, which is easy to solve iteratively. A computer program, developed for this study to implement the MCS method, used an initial value of A from the mean and sigma method to find the value of B. After obtaining a temporary estimate of B, the Newton-Raphson method was used to find a subsequent value of A. The updated value of A was then used to find a new value of B, and so on. The iteration was repeated until a prespecified criterion for the differences for the values of A and B between two successive iterations, was met. For the present study, this criterion was set at .01.