DOCUMENT RESUME

ED 386 466 TM 023 825

AUTHOR	Sykes, Robert C.; Ito, Kyoko
TITLE	The Estimation of Item Difficulty from Restricted CAT Calibration Samples.
PUB DATE	Apr 95
NOTE	27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTCRS	*Adaptive Testing; *Computer Assisted Testing; *Difficulty Level; Estimation (Mathematics); Factor Analysis; Licensing Examinations (Professions); Multiple Choice Tests; *Sample Size; Test Format; *Test Items
IDENTIFIERS	*Calibration; *Dimensionality (Tests)

ABSTRACT

Whether the presence of bidimensionality has any effect on the adaptive recalibration of test items was studied through live-data simulation of computer adaptive testing (CAT) forms. The scurce data were examinee responses to the 298 scored multiple choice items of a licensure examination in a health care profession. Three 75-item part-forms, differing in the degree of bidimensionality, were constructed from the examination forms. The dimensionality of each item was determ:ned based on principal factor loadings using W. F. Stout's (1987) procedure. Samples of 2,100 examinees from an U.S.-educated group and a predominantly foreign-educated group were used, with samples sizes of 100, 200, 400, and 1,000 used for the analyses. As expected, an increase in sample size resulted in greater agreement between calibration and bank b-values. It appeared that dimensionality affected adaptive recalibration. Results from this study varied depending on the dimensionality of the part-forms. Results also affirm the importance of a well-defined reference group for recalibration, since samples from varying ability ranges sometimes produced considerably different b-values. Three tables present details of the analyses. (Contains six references.) (SLD)

The Estimation of Item Difficulty from Restricted CAT Calibration Samples

Robert C. Sykes

Kyoko Ito

CTB/McGraw-Hill

This paper was presented at the Annual Meeting of the National Council on Measurement in Education in San Francisco, April 1995.

INTRODUCTION

Programs employing CAT may require periodic recalibrations of *scored* items for purposes such as monitoring, and possibly correcting for, the effects of scale drift. New recalibration samples that are representative of a reference or total testing population are more readily obtained for paper-and-pencil examinations than for CAT tests. Representative samples may only be obtained in CAT by non-adaptively "seeding" to-be-recalibrated items among adaptively administered (CAT) items. In contrast, samples for CAT items, though readily available, will be non-representative, restricted in range and possibly in size because of the targeting of items to ability.

If item b-values obtained from the one-parameter or Rasch model are sample invariant, the restricted-range samples could be used for purposes of recalibration and other analyses utilizing item parameter estimates (e.g., DIF, model fit). Recalibration utilizing restricted-range CAT samples (i.e., on-line adaptive recalibration) would be more efficient and less costly because all items could be administered adaptively, as opposed to some items adaptively and others nonadaptively.

Previous research based on simulated CAT forms constructed from a form of a paper-and-pencil licensure examination demonstrated that bank b-values were not well replicated when difficult or easy items were recalibrated using responses from more and less able examinees, respectively (CTB/McGraw-Hill, 1993; Ito & Sykes, 1994). The previous research suggested that a "modified" on-line adaptive recalibration might still be a possibility if restricted but larger recalibration samples, similar in mean ability to the reference group, were used for forms that were not too homogenous in item difficulty.

These results, however, did not explicitly control for dimensionality, and because the paper-and-pencil licensure examination that was evaluated has been documented to be bidimensional, may have been impacted by the presence of multidimensionality. The purpose of this study was to evaluate, through live-data simulations of CAT forms, whether the presence of bi-dimensionality has any effect on adaptive recalibration. In this study, test length and the variability of item difficulties were held constant.

METHOD

Source Data

The source data were examinee responses to the 298 scored multiple choice items from a licensure examination in a health care profession. Items selected for this and other full-length operational forms were screened for fit to the Rasch model. The reliability of the selected full-length form was .87 (KR20). Once part-forms were defined (see below), responses to the items in each part-form were extracted from the source data and analyzed.

Part-Forms

Three 75-item part-forms, differing in the degree of bi-dimensionality, were constructed from the examination form. The part-forms are referred to as the "1st-Factor Pure," "2nd-Factor Pure," and "Bi-Dimensional" forms. There were no items in common between the 1st-Factor Pure and 2nd-Factor Pure part-forms, but the Bi-Dimensional part-form did share some items with the other two part-forms.

(Bi-)dimensionality of each item was determined based on "principal factor" factor loadings obtained from Stout's (1987, 1990) program to determine essential unidimensionality, using a large sample drawn from the reference group for the examination. The Stout procedure allows a determination of the essential unidimensionality of the test by utilizing an 'assessment' set of items chosen on the basis of the items' loadings on a second factor or a content appraisal.

A large number of the items in the 1st-Factor Pure form had high loadings on the first factor and low loadings on the second factor. The items in the 2nd-Factor Pure form had relatively high loadings on the second factor and low loadings on the first factor. The Bi-Dimensional form was constructed such that the mean loadings on the first and second factors were both relatively high and similar. Thus, the dimensionality of the part-forms was determined by the number of items with relatively high loadings on the first or second or both factors.

Subsequent analyses using the Stout program confirmed the dimensionality of each part-form. Table 1 shows the eigenvalues, Stout statistics, and the means and standard deviations of absolute first and second factor loadings on the items in the part-forms.

The first eigenvalues for all three part-forms are not large relative to total part-form communality variances (6.406/21.777 = 29.4%, 13.3%, and 23.9% for 1st-Factor Pure, 2nd-Factor Pure, and Bi-Dimensional part-forms, respectively). However, the first eigenvalues for the 1st-Factor Pure and Bi-Dimensional part-forms (6.406 and 4.846, respectively) constitute a greater proportion of the total communality variance of those part-forms than that accounted for by the first eigenvalue for the complete form from which the part-forms were created (15.00/86.09 = 17.4%). The second eigenvalue for the Bi-Dimensional part-form also represents a second factor that is relatively more potent in that part-form (14.8% of total communality variance) than the second factor in the complete form (5.6% of total complete form communality variance). The 2nd-Factor Pure form, despite its unidimensionality, is not as second-factor dominant as the 1st-Factor Pure form is first-factor dominant.

All three part-forms conformed to the test specifications of the examination and hence were test plan representative. The difficulties of the part-forms were made as comparable as possible in terms of the mean and standard deviation of their Rasch difficulty estimates (i.e., b-values). B-value (and p-value) statistics are given in Table 2. All three part-forms had the same mean b-value (-1.15) and the same mean p-value (.75). The standard deviation of b-values was either .60 or .61 and the standard deviation of p-values was either .10 or .11. The correlation between candidate scores on the 1st-Factor Pure and 2nd-Factor Pure forms was .55. The correlations between scores on the Bi-Dimensional form and the 1st-Factor Pure and 2nd-Factor Pure forms was .77 and .71, respectively.

The full-length form, from which the part-forms were constructed, was more difficult (mean b-value = -0.97) and had a larger standard deviation (.79) of b-values. The smaller standard deviations of the part-form b-values were intentional to simulate CAT tests which tend to be more homogeneous in difficulty. The CAT forms simulated in the previous study (1994) had considerably smaller standard deviations of b-values, ranging between .17 and .34.

<u>Attributes of Samples</u>

Large samples from two **subpopulations** were assessed. The first sample was from the reference group of predominantly white first-time U.S.-educated examinees (i.e. "1st-time U.S."). The second sample was from an ethnic group (hereafter "Ethnic Group") that was predominantly foreign-educated.

The mean theta estimates for the representative samples of the 1st-time U.S. and the Ethnic Group examinees (N = 2,100 for each sample) were 0.07 and -0.81, respectively. In the past the Ethnic Group has been one of two groups that had the largest number of items flagged for DIF. Consistently 15% to 18% of the items in the examination forms demonstrate DIF against the Ethnic Group relative to a white reference group.

The part-forms were recalibrated on samples of 1st-time U.S. and Ethnic Group examinees chosen from **four ability (theta) ranges**:

Ability Range	Definition
(1) Restricted 1	-1.0 through -0.5 logits
(2) Restricted 2	-0.5 through 0.0
(3) Full	Unrestricted
(4) Far	As far away from the -1.0 - 0.0 range as possible while still containing at least 800 cases used to create two 400-case samples.

The first three ranges (Restricted 1, Restricted 2, and Full) were common to both subpopulations. The fourth range (Far) was specific to a subpopulation, as shown in the table below:

	Far Range (logits)
1st-time U.S.	From +0.792 to +1.264
Ethnic Group	From -1.578 to -1.074

The Far ranges reflected the groups' relative overall performance levels; that is, the Far range for the reference group contained substantially more able examinees than the Far range for the

Ethnic Group.

Four sample sizes were considered: 100, 200, 400, and 1,000. Except for N = 1,000 and whenever possible, three samples of the same size were obtained. In some cases, only two samples were produced due to insufficient case counts. Samples of 100, 200, and 400 were mutually exclusive. Samples of 1,000 were constructed by pooling samples of smaller sizes.

ANALYSES

The agreement between new calibration b-values and the b-values obtained after the operational administration of the full-length form (i.e. bank b-values) was evaluated. After responses to a given part-form were extracted from the source data for a given sample in a given ability range, the items in the part-form were recalibrated to obtain a new set of one-parameter b-values. The new b-values were then equated to the bank scale so that the mean of the new b-values would be equal to the mean of the corresponding bank b-values. Agreement between new b-values and bank b-values was assessed with two statistics: product-moment correlation (r) and the mean absolute difference (MAD) between the bank b-values and new b-values. B-values for the part-forms were obtained in the same manner as the bank b-values. Maximum likelihood estimates were generated using PARMATE (Burkett, 1991) with item discrimination parameters fixed at 1/D (1/1.7 = .58).

The items in the full-length form had been estimated using responses from a calibration sample of 1,000 1st-time U.S.

examinees, and had been equated to the bank scale so that they would have the same mean as the mean of the b-values obtained from their last previous paper-and-pencil administration. The correlation between the equated (bank) b-values and those obtained from the last previous administrations of the items was 1.00. The MAD for the two sets of b-values was .04.

RESULTS

Because the analysis generated a large number of correlations and MADs, the discussion of b-value agreement will be limited to results that are *averaged* over samples of the same size. Averaged results are indicated in bold face in Table 3.

Agreement between calibration and bank b-values

Table 3 presents the results regarding the b-value agreement. The table, which spans six pages, is arranged first by the group (i.e., 1st-time U.S. vs. Ethnic Group) and by the form (i.e., 1st-Factor Pure, 2nd-Factor Pure, and Bi-Dimensional). The top of the table on each page, just below the title, indicates the group and form to which the page pertains. Each page is then arranged by the ability range (columns; Full, Restricted 1, Restricted 2, and Far Ranges) and by the sample size (rows; 100, 200, 400, and 1000).

1. **Effect of sample size**

As sample size increased, correlations tended to increase and MADs tended to decrease across the 24 combinations of subpopulations (2), part-forms (3), and ability ranges (4). For the 1st-time U.S. sample, correlations monotonically increased between sample sizes of 100 to 1000 for all 12 comparisons and MADs monotonically decreased for all but the Far ability range on the Bi-Dimensional part-form. The correlations did not demonstrate as frequent monotonic increases nor MADs as frequent monotonic decreases for the Ethnic Group.

However, Ethnic Group correlations for the 1000 candidate sample were larger than the average correlations for the 100 candidate samples for all but one of the 12 comparisons (Far ability range, Bi-Dimensional form). MADs for the largest sample were also smaller than the average MADs for the smallest samples (n = 100) for all but two comparisons (Restricted 1 and Far ability ranges for the Bi-Dimensional part-form).

1(a).**Comparisons between sample sizes of 400 and 1000**

As summarized in the table below, the absolute valued differences in correlations and MADs between N = 400 (averaged over two/three N = 400 samples) and N = 1000 were relatively small.

Subpopulation/ Part-Form	Differences (absolute value) in statistics for samples of size 400 & 1,000			
	Correlation		MAD	
	Min.	Max.	Min.	Max.
1st-time U.S.				
1st-Factor Pure	.006	.013	.013	.026
2nd-Factor Pure	.012	.018	.015	.033
Bi-Dimensional	.003	.014	.002	.030
Ethnic Group				
1st-Factor Pure	.001	.006	.005	.012
2nd-Factor Pure	.001	.007	.003	.012
Bi-Dimensional	.000	.012	.002	.011

2. **Comparison among the four ability ranges**

The correlations and MADs were compared to see if there was any consistent pattern, irrespective of subpopulation, form, or sample size. Because the 1st-time U.S. and Ethnic Group subpopulations displayed slightly different tendencies, the subpopulations are discussed separately below.

<u>1st-time U.S.</u> : The Full and Restricted 2 ranges produced similar correlations and MADs, regardless of sample size and part-form. This came as no surprise because the mean of the total group of 1st-time U.S. candidates fell just outside the Restricted 2 range (0.06). All the correlations from these two ability ranges were in the .90s, and the MAD's varied between .091 (1st-Factor Pure, Restricted 2, N = 1000) and .238 (Bi-Dimensional, Full, N = 1000).

The Far ability range yielded b-values that were least similar to bank b-values. The correlations were in the .60s to .80s, and the MAD's ranged from .294 (2nd-Factor Pure, N = 1000) to .516 (Bi-Dimensional, N = 100).

Ethnic Group : The 1st-Factor Pure form yielded results that differed from those from the other two forms. With the 1st-Factor Pure form, both the correlations and MAD's demonstrated that b-values from the Restricted 2 range were consistently the most similar to bank b-values. The correlations for the Restricted 2 range on the 1st-Factor Pure form ranged from .767 to .814, and the MADs from .385 to .435. On the same part-form, the difference in average correlations between the Restricted 2 range and the range that produced the next highest correlation varied between .035 and .074 across four sample sizes.

On the 2nd-Factor Pure and Bi-Dimensional forms, the Ethnic Group candidates from the Restricted 2 range still tended to produce b-values that correlated the best with bank b-values (with the sole exception of the samples of 100 from the Restricted 2 range on the 2nd-Factor Pure part-form). The correlations ranged from .428 to .602. However, the differences in average correlation between the Restricted 2 range and the range that produced the second highest correlation were relatively small, ranging from .009 to .037.

Although the Restricted 2 range tended to produce the greatest correlations, the MADs from this ability range were consistently the greatest on the 2nd-Factor Pure and Bi-Dimensional forms. For instance, the table below shows the mean correlations and MAD's for N = 400.

Part-Form	Ability Range							
	Full		Rest. 1		Rest. 2		Far	
	r	MAD	*r*	MAD	*r*	MAD	*r*	MAD
2nd-Factor Pure	.586	.467	.573	.507	.598	.522	.536	.509
Bi-Dimensional	.405	.587	.382	628	.440	.639	.357	.630

On both the 2nd-Factor Pure and Bi-Dimensional forms for these sample sizes, the Restricted 2 range for the Ethnic Group had the highest correlations <u>and</u> the highest MADs.

3. **Comparison among the three part-forms**

The three part-forms were compared with regard to b-value agreement to see if any consistent pattern would emerge, regardless of subpopulation, ability range, or sample size. With the 1st-time U.S. group, the 1st-Factor Pure form tended to generate b-values the most similar to bank b-values, and the Bi-Dimensional form the least similar. However, as shown in the table below of average correlations for N = 400 and 1000, the differences among the three part forms were relatively small.

	N = 400				N = 1000			
	Ability Range				Ability Range			
Part-Form	Full	Rest. 1	Rest. 2	Far	Full	Rest. 1	Rest. 2	Far
1st-Factor Pure	.972	.939	.971	.885	.985	.945	.982	.896
2nd-Factor Pure	.964	.925	.961	.846	.977	.937	.974	.864
Bi-Dimensional	.964	.883	.956	.782	.976	.891	.970	.785

With the Ethnic Group, the pattern of correlations among the three part forms was considerably more pronounced (see the table below). As with the 1st-time U.S. group, the 1st-Factor Pure part-form produced b-values that were clearly the most comparable to bank b-values, and the Bi-Dimensional form the least comparable. Most of the correlations on the 1st-Factor Pure form were in the .60s and .70s, while those on the 2nd-Factor Pure form were in the .50s (except one, .602, for N = 1000) and those on the Bi-Dimensional form were in the .30s and .40s.

	N = 400				N = 1000			
	Ability Range				Ability Range			
Part-Form	Full	Rest. 1	Rest. 2	Far	Full	Rest. 1	Rest. 2	Far
1st-Factor Pure	.736	.736	.810	.646	.740	.737	.814	.652
2nd-Factor Pure	.586	.573	.598	.536	.593	.572	.602	.539
Bi-Dimensional	.405	.382	.440	.357	.417	.386	.440	.362

With only one exception (1st-time U.S.:1st-Factor Pure part-form:Full ability range for samples of size 200), MADs for the 32 ability-range-by-sample-size-by-subpopulation combinations were larger for the Bi-Dimensional part-form than for the 1st-Factor Pure and 2nd-Factor Pure part-forms. More specific comparisons across the three part-forms revealed that MADs often increased for the Ethnic Group from 1st-Factor Pure (in the .40s) to 2nd-Factor Pure (in the .50s) to Bi-Dimensional part-forms (in the .60s). For the 1st-time U.S. candidates, MADs monotonically increased across the three part-forms in seven out of the eight comparisons involving the larger sample sizes of 400 and 1000. For the Ethnic Group, MADs monotonically increased across the part-forms in all 16 ability-range-by-sample-size comparisons.

CONCLUSIONS

This study investigated the effects of sample size, range of candidate ability, and dimensionality on the estimation of one-parameter b-values. As expected, an increase in sample size resulted in greater agreement between calibration and bank b-values. B-value agreement with a sample size of 400 was comparable to b-value agreement with a sample size of 1,000. Greater differences in b-values were found between samples of 200 and 1000, and sampling fluctuation in results from samples of size 100 was markedly noticeable.

Comparisons among the four ability ranges indicated that for 1st-time U.S. candidates, the Full and Restricted 2 ranges yielded best b-value agreement. For the Ethnic Group candidates, the Restricted 2 range demonstrated best agreement. The Restricted 2 range was closest to the mean ability of the reference group of 1st-time U.S. candidates.

It appears that dimensionality affects adaptive recalibration. Results from this study varied depending on the dimensionality of part-forms. When the three part-forms were compared, both groups produced b-values for the 1st-Factor Pure form that were most similar to the benchmark b-values. Differences among the three part-forms were smaller for the 1st-time U.S. sample.

Specifically, the range in mean correlations across part-forms for the largest 1st-time U.S. sample varied between .009 and .111 when ability range is controlled. The range in mean correlations across part-forms at each of the four ability groups varied between .290 and .374 for the corresponding largest sample of the Ethnic Group. Moreover, MADs from the comparisons involving the Bi-Dimensional part-forms were larger for both subpopulations than those obtained from the comparisons involving the 1st-Factor Pure and 2nd-Factor Pure part-forms.

Results from the study affirm the importance of a well-defined reference group for recalibration. Samples from varying ability ranges produced sometimes considerably different b-values. Moreover, b-values from two subpopulations that differed in educational or training background were often even more disparate. The effects of differences in ability and educational or training background might be expected to confound the estimation of b-values in CAT programs that utilize reference groups that are not relatively homogenous in these characteristics.

References

Burket, G.R. (1990). *PARMATE Version 0.98* [Computer program]. Monterey, CA: CTB/McGraw-Hill.

CTB McGraw-Hill. (February, 1993). Investigations of the stability of NCLEX b-values and the Mantel-Haenszel delta statistic: Implications for CAT. Report submitted to the National Council of State Boards of Nursing.

Ito, K. & Sykes, R.C. (April, 1994). The effect of restricting ability distributions in the estimation of item difficulties: Implications for a CAT implementation. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 52, 589-618.

Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 298-325.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.

Table 1

Results from the Stout Analyses of Three Part-Forms[1]

	Part-Forms					
	1st-Factor Pure		2nd-Factor Pure		Bi-Dimensional	
Eigenvalues:	Eigen.	Diff.	Eigen.	Diff.	Eigen.	Diff.
	6.406	4.532	2.219	.433	4.846	1.844
	1.874	.030	1.786	.166	3.003	1.443
	1.843	.215	1.619	.126	1.559	.105
	1.628	.099	1.494	.099	1.454	.074
	1.529	.123	1.394	.032	1.380	.076
	1.406	.023	1.362	.057	1.304	.038
	1.383	.157	1.305	.069	1.266	.041
Communality	21.777		16.654		20.256	
Stout Results						
Stout T:	.049		1.466		2.483	
Prob.	.480		.071		.007	
At α=.05	Unidimensional		Unidimensional		Not unidim.	
Abs. 1st-Factor Loading:						
Mean	.279		.128		.221	
SD	.062		.071		.116	
Abs. 2nd-Factor Loading:						
Mean	.053		.143		.174	
SD	.037		.070		.085	

[1] Based on 1st-time, U.S.-educated candidates.

Table 2

Relative Difficulty of Three Part-Forms[1]

	Part-Forms		
	1st-Factor Pure	2nd-Factor Pure	Bi-Dimensional
Rasch b-Values:			
Mean	-1.15	-1.15	-1.15
SD	.60	.60	.61
Min.	-2.51	-2.84	-2.84
Max.	-.06	.10	.63
P-Values:			
Mean	.75	.75	.75
SD	.10	.10	.11
		Correlations	
2nd-Factor Pure	.55		
Bi-Dimensional	.77	.71	

[1] Based on 1st-time, U.S.-educated candidates.

Table 3

Agreement Between Calibration and Bank B-Values
Subpopulation : First-Time U.S.-Educated
Form : 1st-Factor Pure Form

	Ability Range (θ logits)							
	Full		Restricted 1 (-1.0 - -0.5)		Restricted 2 (-0.5 - +0.0)		Far (+0.792 -+1.264)	
Sample Size	r	MAD	r	MAD	r	MAD	r	MAD
100-1	.884	.250	.925	.205	.918	.201	.744	.505
100-2	.918	.203	.893	.249	.911	.201	.746	.453
100-3	.902	.246					.801	.503
Mean	**.901**	**.233**	**.909**	**.227**	**.915**	**.201**	**.764**	**.487**
200-1	.955	.173	.914	.203	.950	.159	.808	.382
200-2	.951	.169	.929	.176	.941	.171	.823	.386
200-3	.950	.157					.846	.392
Mean	**.952**	**.166**	**.922**	**.190**	**.946**	**.165**	**.826**	**.387**
400-1	.976	.125	.941	.177	.979	.097	.890	.346
400-2	.968	.116	.936	.180	.963	.132	.888	.289
400-3	.972	.119					.876	.291
Mean	**.972**	**.120**	**.939**	**.179**	**.971**	**.115**	**.885**	**.309**
1000	**.985**	**.094**	**.945**	**.166**	**.982**	**.091**	**.896**	**.295**

(continued)

Table 3 (continued)

Agreement Between Calibration and Bank B-Values
Subpopulation : First-Time U.S.-Educated
Form : 2nd-Factor Pure Form

	Ability Range (θ logits)							
	Full		Restricted 1 (-1.0 - -0.5)		Restricted 2 (-0.5 - +0.0)		Far (+0.792 - +1.264)	
Sample Size	r	MAD	r	MAD	r	MAD	r	MAD
100-1	.899	.246	.885	.227	.914	.200	.803	.347
100-2	.918	.211	.845	.268	.906	.201	.810	.382
100-3	.906	.227					.781	.406
Mean	**.908**	**.228**	**.865**	**.248**	**.910**	**.201**	**.798**	**.378**
200-1	.948	.173	.907	.205	.946	.171	.829	.341
200-2	.935	.179	.916	.185	.943	.156	.816	.340
200-3	.947	.172					.794	.404
Mean	**.943**	**.175**	**.912**	**.195**	**.945**	**.164**	**.813**	**.362**
400-1	.970	.118	.926	.184	.965	.127	.846	.316
400-2	.961	.129	.923	.190	.957	.145	.868	.297
400-3	.961	.136					.824	.367
Mean	**.964**	**.128**	**.925**	**.187**	**.961**	**.136**	**.846**	**.327**
1000	**.977**	**.101**	**.937**	**.172**	**.974**	**.112**	**.864**	**.294**

(continued)

Table 3 (continued)

Agreement Between Calibration and Bank B-Values
Subpopulation : First-Time U.S.-Educated
Form : Bi-Dimensional Form

	Ability Range (θ logits)							
	Full		Restricted 1 (-1.0 - -0.5)		Restricted 2 (-0.5 - +0.0)		Far (+0.792 - +1.264)	
Sample Size	r	MAD	r	MAD	r	MAD	r	MAD
100-1	.905	.231	.839	.297	.918	.205	.641	.532
100-2	.906	.232	.816	.284	.895	.220	.683	.528
100-3	.890	.251					.734	.487
Mean	**.900**	**.238**	**.828**	**.291**	**.907**	**.213**	**.686**	**.516**
200-1	.949	.177	.863	.267	.941	.175	.749	.476
200-2	.936	.178	.861	.256	.927	.168	.762	.455
200-3	.958	.148					.723	.489
Mean	**.948**	**.168**	**.862**	**.262**	**.934**	**.172**	**.745**	**.473**
400-1	.968	.128	.882	.220	.960	.142	.780	.458
400-2	.959	.139	.883	.255	.951	.155	.780	.445
400-3	.966	.128					.787	.415
Mean	**.964**	**.132**	**.883**	**.238**	**.956**	**.149**	**.782**	**.439**
1000	**.976**	**.110**	**.891**	**.227**	**.970**	**.119**	**.785**	**.441**

(continued)

Table 3 (continued)

Agreement Between Calibration and Bank B-Values
Subpopulation : Ethnic Group
Form : 1st-Factor Pure Form

	Ability Range (θ logits)							
	Full		Restricted 1 (-1.0 - -0.5)		Restricted 2 (-0.5 - +0.0)		Far (-1.578 - -1.074)	
Sample Size	r	MAD	r	MAD	r	MAD	r	MAD
100-1	.670	.496	.742	.521	.756	.439	.635	.478
100-2	.758	.421	.721	.464	.778	.430	.643	.483
100-3	.721	.434					.639	.490
Mean	**.716**	**.450**	**.732**	**.493**	**.767**	**.435**	**.639**	**.484**
200-1	.735	.426	.719	.471	.795	.391	.645	.460
200-2	.736	.421	.754	.431	.766	.441	.648	.492
200-3	.710	.429					.661	.450
Mean	**.727**	**.425**	**.737**	**.451**	**.781**	**.416**	**.651**	**.467**
400-1	.734	.426	.742	.454	.813	.390	.636	.469
400-2	.733	.425	.730	.459	.806	.393	.658	.483
400-3	.741	.424					.644	.478
Mean	**.736**	**.425**	**.736**	**.457**	**.810**	**.392**	**.646**	**.477**
1000	**.740**	**.417**	**.737**	**.452**	**.814**	**.385**	**.652**	**.466**

(continued)

Table 3 (continued)

Agreement Between Calibration and Bank B-Values
Subpopulation : Ethnic Group
Form : 2nd-Factor Pure Form

	Ability Range (θ logits)							
	Full		Restricted 1 (-1.0 - -0.5)		Restricted 2 (-0.5 - +0.0)		Far (-1.578 - -1.074)	
Sample Size	r	MAD	r	MAD	r	MAD	r	MAD
100-1	.542	.516	.542	.561	.523	.565	.520	.526
100-2	.566	.480	.507	.538	.497	.598	.521	.525
100-3	.501	.565					.500	.520
Mean	**.536**	**.520**	**.525**	**.550**	**.510**	**.582**	**.514**	**.524**
200-1	.585	.479	.551	.540	.596	.514	.523	.530
200-2	.581	.480	.564	.492	.574	.571	.507	.533
200-3	.525	.512					.510	.527
Mean	**.564**	**.490**	**.558**	**.516**	**.585**	**.543**	**.513**	**.530**
400-1	.577	.474	.586	.485	.597	.514	.529	.511
400-2	.597	.471	.560	.528	.598	.529	.546	.499
400-3	.583	.457					.533	.516
Mean	**.586**	**.467**	**.573**	**.507**	**.598**	**.522**	**.536**	**.509**
1000	**.593**	**.464**	**.572**	**.502**	**.602**	**.510**	**.539**	**.502**

(continued)

Table 3 (continued)

Agreement Between Calibration and Bank B-Values
Subpopulation : Ethnic Group
Form : Bi-Dimensional Form

Sample Size	Full r	Full MAD	Restricted 1 (-1.0 - -0.5) r	Restricted 1 (-1.0 - -0.5) MAD	Restricted 2 (-0.5 - +0.0) r	Restricted 2 (-0.5 - +0.0) MAD	Far (-1.578 - -1.074) r	Far (-1.578 - -1.074) MAD
100-1	.379	.621	.374	.633	.429	.648	.356	.633
100-2	.409	.542	.395	.601	.427	.683	.358	.617
100-3	.440	.634					.376	.605
Mean	**.409**	**.599**	**.385**	**.617**	**.428**	**.666**	**.363**	**.618**
200-1	.437	.572	.389	.644	.425	.632	.339	.616
200-2	.403	.582	.403	.590	.445	.675	.331	.644
200-3	.354	.639					.343	.627
Mean	**.398**	**.598**	**.396**	**.617**	**.435**	**.654**	**.338**	**.629**
400-1	.400	.605	.395	.618	.435	.643	.345	.632
400-2	.414	.576	.369	.638	.444	.634	.382	.613
400-3	.401	.580					.344	.644
Mean	**.405**	**.587**	**.382**	**.628**	**.440**	**.639**	**.357**	**.630**
1000	**.417**	**.580**	**.386**	**.626**	**.440**	**.632**	**.362**	**.619**

Ability Range (θ logits)