

True Score Equating by Fixed *b*'s Scaling: A Flexible and Stable Equating Alternative

Marilyn M. Hicks
Educational Testing Service

Six methods of equating The Test of English as a Foreign Language (TOEFL) test scores were evaluated in terms of scale stability. True score item response theory (IRT) equating based on "Fixed *b*'s" scaling, the current TOEFL operational scaling and equating procedure, was found to produce the least discrepant results when compared to two IRT models (*b* parameter estimated, *a* and *c* parameters fixed; all three parameters reestimated), and to three conventional equating methods (Tucker, Levine, and equipercentile). The results for Fixed *b*'s scaling were limited by an inadequately fit item; but if such items can be identified prior to calibration, or if pretested data are observed to produce reliable estimates of total group data, then true score IRT equating based on scaling by fixing the *b* parameters of a set of pretested items may be a very acceptable option.

Existing test equating methodologies are, in general, based on assumptions regarding the population and/or test characteristics which may be difficult to evaluate in a given application. The appropriateness of a method can be assessed in terms of the degree to which the stability of the score scale is maintained over a series of equatings. Such an empirical appraisal may implicitly validate the tenability of the assumptions to the data or may be an indicator of the robustness of the method, given the data. In any case, adoption of an equating methodology which produces minimal scale drift over

a chain of equatings is critical to accurate measurement.

IRT Scaling and Equating of TOEFL Test Scores

The Test of English as a Foreign Language (TOEFL), which assesses the English proficiency of foreign students desiring to study at colleges and universities in the United States and Canada, is comprised of three sections: Section I, Listening Comprehension; Section II, Structure and Written Expression; and Section III, Reading Comprehension and Vocabulary. An equated score is reported for each section in addition to a total score as a weighted sum of the equated section scores. In September 1978 TOEFL adopted item response theory (IRT) methodology in the form of the three-parameter logistic model for the purpose of equating in lieu of conventional linear methods. In this case, the three-parameter logistic model for item *i*,

$$P_i = P_i(\theta) = c_i + (1 - c_i) \times \{1 + \exp[-1.7a_i(\theta - b_i)]\}^{-1}, \quad [1]$$

requires the estimation of three item parameters, *a*, *b*, and *c*, and an ability parameter, θ . A measure of the discriminating power of the item, the *a* parameter, is related to the slope of the item curve $P_i(\theta)$ at the point of inflection. The *b* parameter is that value on the ability scale midway between the upper and lower asymptotes of the logistic item curve. As a location parameter, it is an index of

255

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 3, Summer 1983, pp. 255-266
© Copyright 1983 Applied Psychological Measurement Inc.
0146-6216/83/030255-12\$1.85

the item difficulty. The c parameter is the value of the ordinate at the lower asymptote of the item curve and is associated with the effects of guessing on the item.

Using LOGIST (Wood, Wingersky, & Lord, 1976), TOEFL parameters are estimated such that θ is scaled to mean zero and standard deviation one, with b 's on the θ scale. If another group of examinees were administered the same item and a similar scaling were applied, any differences in level and spread of ability between the two groups would result in dissimilar values of the b 's. The invariance of item parameters across groups and θ estimates across tests will hold only if parameter estimates derived from subsequent groups are placed on some established scale. If a set of items have been scaled on a given group of examinees, estimates based on successive groups can be linearly transformed to the established scale. When old and new forms are linked by a block of common items, the slope and intercept parameters of the line relating the b 's can be used to scale all the items in the new form (Marco, 1977). Stocking and Lord (1982) have developed a linear transformation which results from the minimization of the average squared difference between true score estimates and have reported favorable results for this method.

Current TOEFL scaling procedures do not depend on a block of items common to two forms; instead, calibrated (scaled) pretested items, selected from many previous test forms, serve as the equating items in each version of the test. During parameter estimation, the a and c parameters for the calibrated items are reestimated, but the b parameters are held fixed at the values derived in the initial calibrations. Alluded to as "Fixed b 's" scaling, the presence of the precalibrated items sets the scale for the noncalibrated items. In common item equating, the equating items are selected to be representative of the total test in content and other specifications; however, the precalibrated items in the Fixed b 's scaling need only be chosen to span the range of difficulty and discriminating power of the total test. Implicit in this procedure is the basic IRT assumption that the estimates of difficulties will hold for all testing groups except for scale

factors. Plots of item characteristic curves, on which were superimposed squares representing the observed proportions of examinees at a given ability level responding correctly to the items (item ability regressions) indicated that, on occasion, some of the precalibrated items did not adequately reflect the response patterns of the current examinee group. The fit of the newly calibrated items was usually quite satisfactory.

Once the item parameters are on scale, it is only necessary to calculate the sum of the item characteristic curves, the test characteristic function, which specifies true scores as a function of ability. Two scores on the tests are then considered equivalent if they depend on the same value of θ (Lord, 1980, pp. 199–205).

Conventional Equating Methods of This Study

Conventional equating methods of this study included linear and equipercenile equating, which are defined as follows:

Equipercenile equating: For a given group of examinees, two scores on separate forms of a test are considered equivalent if their percenile ranks are equal.

Linear equating: For a given group of examinees, two scores on separate forms of a test are considered equivalent if they correspond to equal standard score deviates. (Angoff, 1982)

In the usual testing situation, where separate groups take the two test forms, the strategy utilized in implementing these definitions involves the formation of a synthetic equating population, T , as a weighted composite of the two testing groups— P , the group taking the new form and Q , the group taking the old form—such that

$$T = w_1P + w_2Q, \quad [2]$$

where w_1 and w_2 are weights assigned to the two groups. In the case of common item equating, information derived from an anchor test, a set of items common to both forms, aids in determining the distributions and first two moments of the synthetic equating group. Details of these procedures

are given in Braun and Holland (1982) and Angoff (1982).

Method

Equating Design

Each of the seven experimental administrations of this study was comprised of several subtests which included both operational and nonoperational items. Each link in the experimental chain was formed by the inclusion of operational items from the previous form as shown below (these items were nonoperational in the current form):

Form	Operational	Pretest Slots
T1	a, a*	
T2	b	a
T3	c	b
T4	d	c
T5	e	d
T6	f	e
T7	g	f
T8 (chain)		g
T8 (direct)		a*

For some of the equating models, this resulted in anchor tests that were internal to the old form and external to the new form. For Sections II and III six types of equating were included in this study as follows:

1. Modified Three Parameter: a and c parameters were held fixed at values determined to be representative of current TOEFL data; only the b 's were reestimated. For Section II, a was fixed at 1.00 and c at .19. For Section III, the fixed value of a was 1.03 and c was .20. Parameters were scaled using the Stocking and Lord (1982) characteristic curve transformation.
2. Three Parameters Reestimated: All three parameters were reestimated and scaled using the characteristic curve transformation. No upper bound on estimates of the a parameter.
3. Fixed b 's IRT: This replicated the current TOEFL operational scaling procedures as pre-

viously described; b 's were held fixed at pretested values, only a and c parameters were reestimated. As a result, the equating items were a different set than those used in the linked forms equating and were internal to the test. An upper limit of 1.5 was placed on the estimates of the a parameter; thus, $0 < a \leq 1.5$.

4. Tucker Linear Equating: Tucker parameters were used throughout the chain of equatings.
5. Levine Linear Equating: Levine parameters were used throughout the equating chain.
6. Equipercentile Equating.

For each IRT and conventional equating condition, a separate base form (T1) scale was established. For all IRT equatings, the experimental form was equated to the appropriate version of the base form. The links served only for the purpose of scaling in the Modified Three Parameter and Three Parameters Reestimated models; while in the three conventional equating methods, each experimental subtest was equated to the previous form in the chain. The equating group for the Fixed b 's method was a spaced sample across all subtests of the experimental forms. All other equating groups were necessarily based on the single subtest which served as the link. IRT equatings were derived from operational TOEFL computer programs. Tucker, Levine, and equipercentile equatings were generated through standard programs used at Educational Testing Service.

Data Analysis

The design accounted for an empirical evaluation of the stability of the various equating methods by utilizing two subtests of the final experimental form, T8. Accordingly, the following items were included in these subtests:

1. In T8 (ch) a set of items linked to the previous form in the equating chain.
2. In T8 (dir) a set of items from T1 as a direct link to the base form.

The equatings derived from the direct link served as the criterion against which each equating chain would be compared using a discrepancy index de-

veloped by Petersen (Petersen, Marco, & Stewart, 1982) and a computer program written by staff in College Board Statistical Analysis. The index is a weighted mean square difference decomposed into the variance of the difference and the squared bias. Thus, if $d_i = (t'_i - t_i)$, where for raw score i , $i = 0, 1, \dots, n$, t'_i and t_i are converted scores corresponding to the criterion and chain equatings, respectively, and f_i is the number of examinees at each score level, then

$$\sum f_i d_i^2 / n = \sum f_i (d_i - \bar{d})^2 / n + \bar{d}^2, \quad [3]$$

i.e., Total Error Squared = Variance of Difference + Squared Bias.

Optimum conditions for the criterion comparisons include equivalent samples and anchor tests of equal difficulty for the two subtests. All equating comparisons were based on independent samples taking the two T8 experimental subtest forms. Comparisons involving equipercentile equatings were limited to the range of scores actually observed.

Results

Characteristics of the Forms Used in the Study

Mean equated deltas (transformed item difficulties, see Angoff & Dyer, 1971) for the forms used in the equating comparisons are given in Table 1. The range of delta values observed for TOEFL is 4 to 18; middle difficulty is 11.7. The character-

istics of the forms used in the equating comparisons closely parallel Variation 8 in the Petersen et al. (1982) study in that, for some equatings, the base form was slightly more difficult than the test to be equated and, for Section III of T8 (dir), the anchor test was more difficult than the operational test. These conditions were found to rather consistently produce greatest error in the evaluation of linear equating (Petersen et al., 1982, Table 10). Raw score data for the equating samples are given in Table 2, where slightly higher means were observed for T8 (ch).

The results described may have implications for the equating comparisons in Section III. A common procedure in evaluating the results of an equating experiment has been the use of the identity equating (Levine, 1955; Petersen et al., 1982). In this case, the base form is readministered as the final link, and lack of scale stability is evaluated in terms of the departure of the slope of the equating line from unity. Objections to this method involve the possible advantage derived from equating a test to itself in the case of the one-parameter IRT model (Petersen et al., 1982). An alternative procedure of using two variants of a form, one based on a direct link to the scale and the other the result of the chain, was adopted in this study to circumvent this objection. Equivalent samples for these forms were assumed to be attainable by virtue of spiralling (distribution of the subtests in serial order). The characteristics of the tests used in the equating comparisons are summarized in Table 3.

Table 1
Mean Equated Deltas for Operational and Anchor Tests
Used in the Equating Comparisons

Form	Section II		Section III	
	Operational	Anchor	Operational	Anchor
T1	11.9	11.9	12.6	12.5
T7	11.8	11.9	12.0	12.3
T8(ch)	11.9	11.8	12.1	12.2
T8(dir)	11.9	12.0	12.1	12.5

Table 2
Raw Score Means, Standard Deviations, and Sample Sizes
for All Equating Groups

Equating Method and Form	N	Section II		Section III	
		Mean	S. D.	Mean	S. D.
Base Form					
T1 (IRT)	14068	23.27	7.01	31.43	10.13
T1 (Conv.)	4580	23.88	7.06	32.03	10.20
Experimental Forms, Fixed <i>b</i> 's Equating					
T2	2283	25.30	6.86	35.77	10.05
T3	1159	25.24	6.68	34.71	10.60
T4	2271	25.45	7.00	36.09	10.15
T5	1774	26.27	6.81	37.10	10.01
T6	2426	25.34	7.09	34.38	9.16
T7	2330	25.38	6.62	37.38	9.65
T8 (ch)	1011	26.35	6.29	38.20	8.66
T8 (dir)	988	26.23	6.97	37.26	8.61
Experimental Forms, Other Equatings*					
T2	1265	26.39	6.98	35.98	10.20
T3	1575	25.98	6.77	35.59	10.48
T4	1530	25.96	6.93	36.59	9.88
T5	1275	26.62	6.68	37.37	9.90
T6	1710	26.14	7.01	35.13	9.20
T7	1825	26.24	6.37	38.64	9.21
T8 (ch)	1005	26.44	6.34	38.23	8.66
T8 (dir)	980	25.95	6.49	37.72	8.71

*Tucker, Levine, Equipercenile, Modified Three-Parameter, Three-Parameter Reestimated.

Discrepancy Indices

Discrepancy indices for the six equating methods are listed in Table 4. Least error was observed for Fixed *b*'s scaling in both sections, with Modified Three Parameter and Tucker equating following in the order of magnitude of error. A positive bias indicates that the criterion tended to produce higher scores than the chain, and conversely for negative bias. In Section II the chain results underestimated the criterion scores; while in Section III the criterion was overestimated, this latter effect probably due, in part, to the variations in difficulty described above. Indeed, the major effect of the variations

observed in Section III was the direction of bias; however, Fixed *b*'s equating was the least sensitive to these differences.

The magnitude of the proportion of squared bias for Modified Three Parameter is observed to be quite large for both sections. Although the error for the Three Parameters Reestimated model was large compared to other IRT methods, most of this error was due to the variance of the differences. These results are inherent in the models, however. The constant values of the *a* parameter in the Modified Three Parameter vary from form to form only by division of the slope of the linear transformation, which limits the range of the slopes of the

Table 3
 Characteristics of Test Forms Used in Equating Comparisons

Section II	Section III
1. Base form and operational test are of equivalent difficulty for all equatings.	1. Base form more difficult than the test to be equated for IRT equatings. For conventional equating, base form and test to be equated were of equal difficulty.
2. Anchor tests of equivalent difficulty for forms in the equating comparisons.	2. Dissimilar difficulty of anchor tests for forms in the equating comparisons.
3. Anchor test roughly equivalent in difficulty to operational test.	3. Anchor test relatively more difficult than operational test.

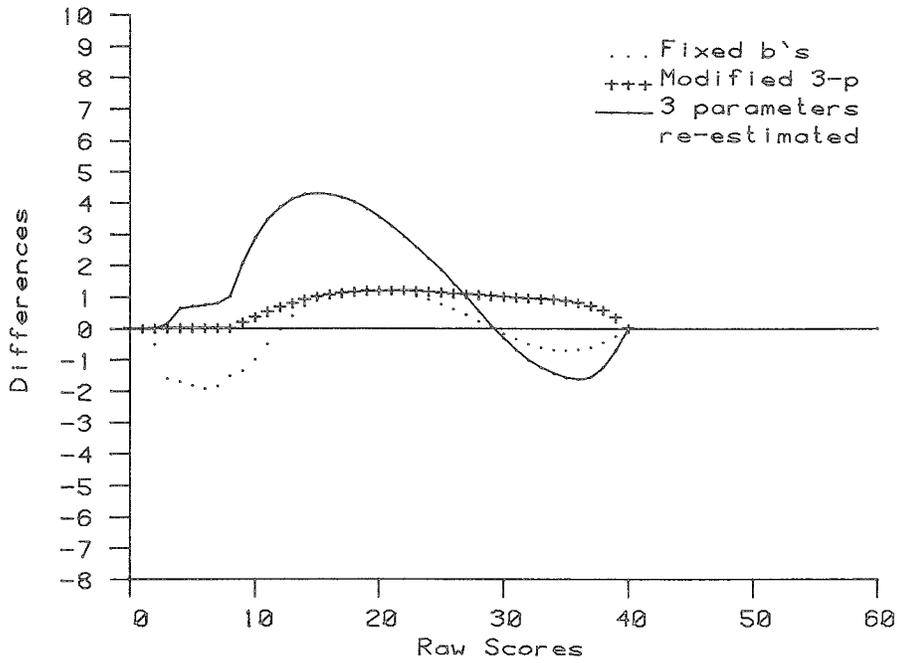
test characteristic curves. When compared to the criterion, the major difference is simply a shift in location. As a result, the variability of the differences will be a small portion of the total error. On the other hand, the slopes of the test characteristic functions for the three-parameter model can vary substantially, accounting for less systematic differences. These effects can be seen in graphs of the unweighted differences between the criterion and chained results in Figures 1 and 2. For linear equating, the graph of the differences is simply a line of negative slope; the greater the absolute value of the slope, the greater the bias.

Standard errors of measurement ranged from 2.92 to 4.03 for Section II and from 2.61 to 3.20 for Section III. Other studies have determined that the standard error of equating is generally less than the standard error of measurement. Equating errors are larger at the tails of the distribution and, among equating methods, largest for equipercenile equating (Lord, 1981a). The mean difference for all criterion comparisons fell within the range of the standard error of measurement. The upper and lower limits of the converted score scale are, in part, determined by the method of equating. For the IRT equatings these limits are the scaled scores at the

Table 4
 Equating Criterion Comparisons

Method	Section II			Section III		
	Var.	Bias	Error	Var.	Bias	Error
Modified						
Three-Parameter	.04	(+)1.04	1.08	.35	(-)1.74	2.09
Fixed b's	.52	(+) .10	.62	.21	(+) .61	.82
Three-Parameter	3.84	(+)1.64	5.48	3.48	(-)2.36	5.84
Tucker	1.38	(+)2.55	3.93	1.10	(-)1.34	2.44
Levine	3.19	(+)4.02	7.21	2.48	(-)2.20	4.68
Equipercenile	2.00	(+)4.61	6.61	.51	(-)4.41	4.92

Figure 1
 Unweighted Differences between Direct and Chain Equatings for Section II
 (a) IRT Equatings



(b) Conventional Equatings

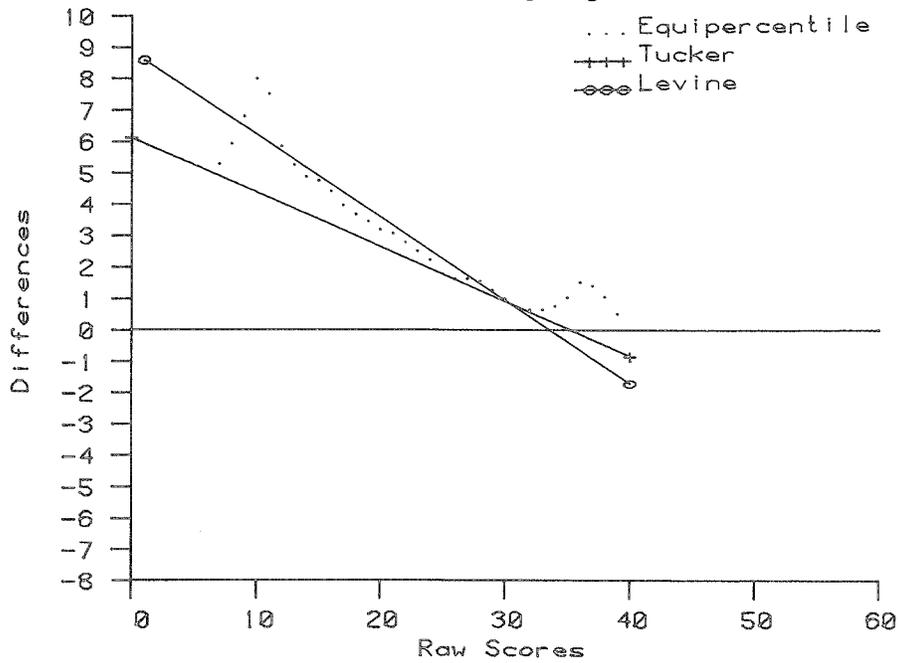
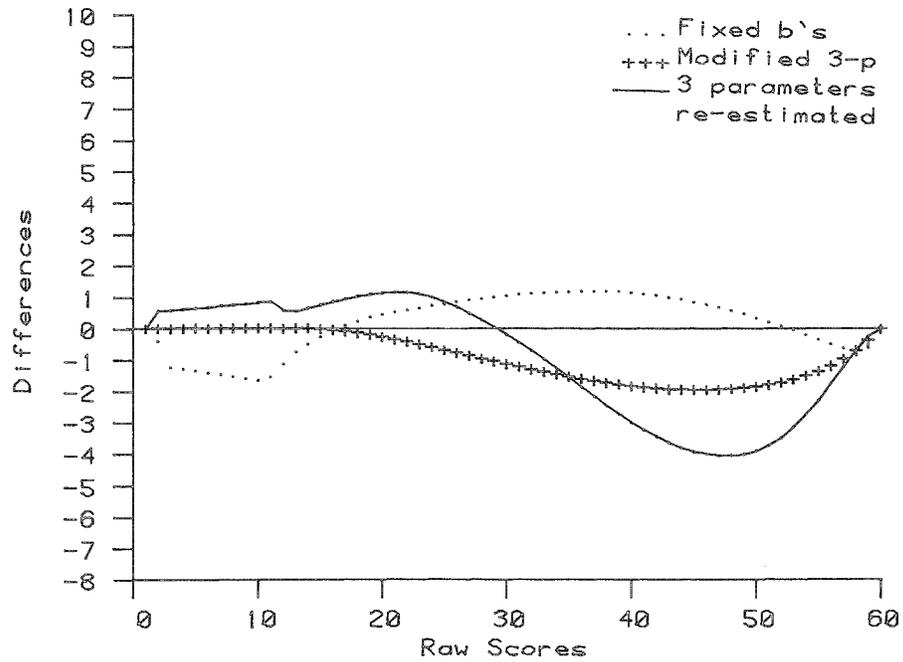
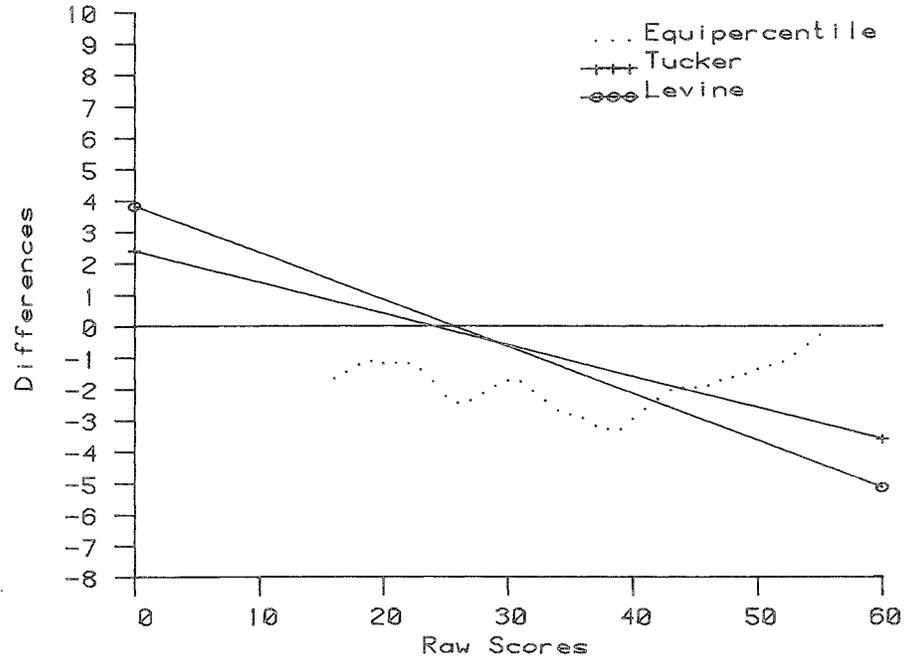


Figure 2
 Unweighted Differences Between Direct and Chain Equatings for Section III
 (a) IRT Equatings



(b) Conventional Equatings



upper asymptote of the test characteristic curve of the old form; for TOEFL, a lower limit of 20. In the equipercentile equatings of this study, the upper and lower limits of the converted scores correspond to the range of observed raw scores. Depending on how the slopes differ in linear equating, greatest differences will generally occur at either or both extremes of the scale.

Methods Comparisons

Discrepancy indices between methods based on the direct and chain results are given in Tables 5 and 6. Differences observed in the two tables are illustrative of a major source of error in equating. From Table 5, all else being equal, the various methods produce comparatively similar results, while discrepancies listed in Table 6 reflect, among other

things, the variability due to methods of linking the forms. From Table 6 can be observed the not too surprising result that Tucker and Levine equatings are the most similar when the effects of linking are taken into account. Among the largest differences observed are the discrepancies between Fixed b 's and the Three Parameters Reestimated, which probably incorporates some of the effects of reestimating the b -parameters versus holding them fixed. Modified Three Parameter versus Fixed b 's have smallest error among all the IRT comparisons. It can also be observed that the values of total error in Section II tend to be higher than those in Section III. This may be due to the fact that a 41-point observed score scale in Section II, as contrasted with a 61-point raw score scale for Section III is being stretched to one that can theoretically range from 20 to 80.

Table 5
Total Error and Squared Bias* Comparisons
of Equating Methods, Direct Results

Section and Method	Method**					
	1	3	F	T	L	E
Section II						
1	---	.27	.44	.26	.07	.29
3	.12	---	.06	.14	.14	.11
F	.12	.00	---	.30	.44	.11
T	.23	.02	.02	---	.08	.28
L	.04	.02	.14	.08	---	.17
E	.03	.04	.03	.11	.00	---
Section III						
1	---	.09	.07	.52	.31	.27
3	.01	---	.07	.96	.58	.26
F	.05	.01	---	.69	.29	.19
T	.04	.10	.18	---	.12	.51
L	.02	.06	.02	.00	---	.30
E	.02	.00	.00	.08	.05	---

*Total error above diagonal, squared bias below diagonal.

**1 = Modified Three-Parameter, 3 = Three Parameters Reestimated, F = Fixed b 's, T = Tucker, L = Levine, E = Equipercentile.

Table 6
Total Error and Squared Bias* Comparisons
of Equating Methods, Chain Results

Section and Method	Method**					
	1	3	F	T	L	E
Section II						
1	---	3.95	1.70	2.57	4.60	3.36
3	.37	---	9.68	1.43	1.33	1.47
F	1.01	2.62	---	7.94	11.06	9.35
T	1.12	.20	4.26	---	.36	.17
L	1.42	.34	4.83	.02	---	.35
E	1.64	.44	5.26	.04	.00	---
Section III						
1	---	1.29	.49	1.26	1.52	1.43
3	.11	---	3.31	1.20	.86	3.66
F	.09	.41	---	2.39	3.09	1.56
T	.13	.48	.00	---	.18	3.79
L	.00	.10	.11	.15	---	3.51
E	.81	.33	1.43	1.53	.73	---

*Total error above diagonal, squared bias below diagonal.

**1 = Modified Three-Parameter, 3 = Three Parameters Reestimated, F = Fixed b 's, T = Tucker, L = Levine, E = Equipercentile.

Discussion and Conclusions

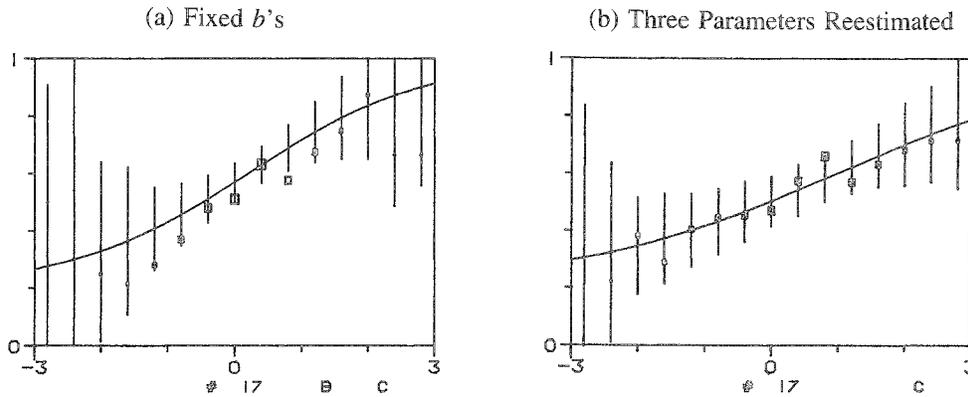
Fixed b 's Scaling

It is not surprising that IRT equating based on Fixed b 's scaling would produce such excellent results in terms of the criterion of this study, since the location parameters for half (or more) of the items in each section are fixed with only the a and c parameters allowed to vary. Assuming that the b parameters held for subsequent groups, bias in the a parameters would be a major source of error. Positive statistical bias does exist for the a 's and is greatest for highly discriminating, difficult items (Lord, 1981b, 1982). In Fixed b 's scaling, an upper limit of 1.5 is placed on the estimated a parameter, which may reduce the effect of bias for this group of items. Plots of precalibrated vs. reestimated a 's collected over time have exhibited no obvious evidence of bias, differing only in degree of scatter

about the line through the origin. A detailed analysis of the precalibrated and reestimated a 's has also failed to detect any evidence of bias. In practical terms, Fixed b 's equating offers flexibility and item security that cannot be derived from methods of equating based on a block of items common to two forms, since compromise of the first form can jeopardize an entire future administration.

As noted earlier, Fixed b 's equating has been observed to occasionally result in poor fit among precalibrated items. An example of an item better fit by reestimating the b parameter is given in Figure 3. This was the most deviant fit of the precalibrated items in these comparisons. Precalibrated items that are identified as seriously aberrant in terms of fit might be treated as noncalibrated, and all parameters reestimated on the current group. Such items could be identified prior to item calibrations by comparing equated deltas based on pre-

Figure 3
Item Ability Regressions for an Item Scaled by Fixed b 's and Three Parameters Reestimated



testing with those derived in a preliminary item analysis (equated deltas and b parameters have been found to correlate very highly, approximately .96). Such a procedure is workable so long as these items remain a small proportion of the precalibrated items.

The possibility that the occasional poor item fit for Fixed b 's scaling is due to sampling error was suggested by results observed when parameters were estimated on an increased sample size, in which case an inadequately fit item based on a sample of approximately 2,000 was considerably improved when the parameters were estimated on a sample of 14,000. Unfortunately, such remedial procedures are quite expensive.

Modified Three Parameter

Results for the Modified Three Parameter method were quite satisfactory for both sections. A practical advantage to this method is the smaller sample size required for parameter estimation, which would have material impact on the difficulties involved in maintaining a precalibrated item pool. Associated with this is the reduction in computer costs for estimating parameters.

Three Parameters Reestimated

The relatively large error associated with estimating all three parameters may reflect the "true"

effects of the variability associated with TOEFL testing groups. Fixed b 's scaling, as implemented by TOEFL, might be categorized as a less sensitive model by virtue of the constraints imposed on the variation of some of the b parameters and the limits on the a parameters. In contrast, a great deal more information about the current group is introduced into the scaling process in estimating all three parameters. One major effect which can be inferred from the graphs of the differences in Figures 1 and 3 is that through the chain the a 's tended to be higher for the more difficult items and lower for the easier items for the Three Parameters Reestimated model. For fixed b 's, the unweighted difference curves indicated that the slopes of the two test characteristic curves were quite comparable (less differences in the a parameters).

Conventional Equating Methods

Of the linear methods, Tucker equating produced the best results, outperforming the Three Parameter Reestimated IRT model. It might be concluded that basic assumptions of Levine equating were not met by the data as, for example, the requirement of parallelism of the anchor and operational tests. The observed differences in the anchor tests did not seem to affect the magnitude of error but did affect the direction of bias.

Conclusions

A critical consideration of IRT equating is the method used to scale the parameters in order to preserve the property of invariance (Lord, 1980, chap. 3). Fixed b 's scaling was observed to produce better results than the linear transformation for TOEFL data in terms of the criterion of this study, but at a price in terms of an inadequately fit item. Although it has been observed that such items tended to be more difficult upon retesting, no systematic evaluation of these differences has been conducted. The deviations from the theoretical curve observed in this study would probably have little or no effect on the conversions; however, with proper monitoring of the data this method of scaling could be quite effective. Furthermore, it offers great flexibility in that the equating items can be selected from many forms instead of linked to a single previous form and need not be constrained to be parallel in terms of all specifications as is the case of common item equating. This may be a useful feature in terms of test construction, providing greater leeway in item selection. As indicated by the results for Section III, Fixed b 's scaling was rather robust under less than ideal conditions.

References

- Angoff, W. H. Summary and derivations of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. New York: Academic Press, 1982.
- Angoff, W. H., & Dyer, H. S. The admissions testing program. In W. H. Angoff (Ed.), *The College Board Admissions Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board, 1971.
- Braun, H. I., & Holland, P. W. Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. New York: Academic Press, 1982.
- Levine, R. S. *Equating the score scales of alternate forms administered to samples of different ability* (RB-55-23). Princeton NJ: Educational Testing Service, 1955.
- Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum, 1980.
- Lord, F. M. *The standard error of equipercentile equating* (RR-81-48). Princeton NJ: Educational Testing Service, 1981.(a)
- Lord, F. M. *Unbiased estimators of ability parameters, of their variance, and of their parallel forms reliability* (RR-81-50). Princeton NJ: Educational Testing Service, 1981.(b)
- Lord, F. M. *Statistical bias in maximum likelihood estimators of item parameters* (RR-82-20-0NR). Princeton NJ: Educational Testing Service, 1982.
- Marco, G. L. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 1977, 14, 139-160.
- Petersen, N. S., Marco, G. L., & Stewart, E. E. A test of the adequacy of linear score equating methods. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. New York: Academic Press, 1982.
- Stocking, M. L., & Lord, F. M. *Developing a common metric in item response theory* (RR-82-25-0NR). Princeton NJ: Educational Testing Service, 1982.
- Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOG-IST: A computer program for estimating examinee ability and item characteristic curve parameters* (RM-76-6). Princeton NJ: Educational Testing Service, 1976.

Acknowledgments

The research reported here was supported by the TOEFL Research Committee. Appreciation is extended to Ronice Morgan and Dawn Robinson for their technical assistance.

Author's Address

Send requests for reprints or further information to Marilyn M. Hicks, Educational Testing Service, Princeton NJ 08541, U.S.A.