Standard Error of an Equating by Item Response Theory

Frederic M. Lord
Educational Testing Service

A formula is derived for the asymptotic standard error of a true-score equating by item response theory. The equating method is applicable when the two tests to be equated are administered to different groups along with an anchor test. Numerical standard errors are shown for an actual equating (1) comparing the standard errors of IRT, linear, and equipercentile methods and (2) illustrating the effect of the length of the anchor test on the standard error of the equating.

Some psychometricians currently regard true-score equating by item response theory (IRT) as the method of choice for equating nonparallel forms of unidimensional tests (for example, Cowell, 1979; Petersen, Cook, & Stocking, 1981; Yen, 1982) and are routinely applying this true-score equating to actual observed scores. To date, no information has been available about the sampling error of such equatings. In particular, it has not been obvious, supposing the true relationship to be linear, whether a true-score equating using IRT has larger or smaller sampling error than a conventional linear equating based on means and standard deviations. Neither has it been known how IRT equating compares in this respect to conventional equipercentile equating. The present article is a first step in illuminating these questions. Additionally, the formulas derived here provide a badly needed basis for deciding how long an anchor test must be in order to provide a satisfactory equating link.

In IRT an examinee's expected number-correct score ξ on Test X is equal to the test characteristic function evaluated at the examinee's ability level θ:

$$\xi = \sum_{g=1}^{n_x} P_g(\theta) \qquad [1]$$

where $P_i(\theta)$ is the item response function, the probability of a correct answer to item i at ability level θ. If there is a second Test, Y, measuring the same ability as X, the expected number-correct score η on this test may be written as

$$\eta = \sum_{h=1}^{n_y} P_h(\theta) \ . \qquad [2]$$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 4, Fall 1982, pp. 463–472

0416-6216/82/040463-10$1.50

Equations 1 and 2 are parametric equations for the functional relationship between ξ and η. Note that this relationship is an exact mathematical one, not a statistical association. Given any θ, Equations 1 and 2 determine a pair of values, ξ and η, that represent the same ability level as θ. Pairs of values (ξ, η) determined in this way are *equated.* In practice, it is often assumed that the functional relationship of η and ξ given by Equations 1 and 2 can also be applied to actual number-correct scores on the two tests, producing an equating of these scores.

This article deals only with the sampling errors in estimating the equating relationship of η to ξ. In Equations 1 and 2, estimated item parameters must be used for determining the probability of a correct response to an item. These are the source of the sampling errors in IRT equating. Note that the ability estimates for individual examinees are not used in Equations 1 and 2 and thus will not appear in the formulas. Until now, the sampling errors of IRT equatings have never been estimated.

Data

In IRT equating, there frequently is a set of common items that are administered to all examinees. These are needed in order to get Test Y item parameters on the same scale as Test X item parameters. If the common items are external to Tests X and Y, as assumed here, the common items are called the *anchor test,* or, in the present report, Test W. The sampling variance formulas to be obtained here can be modified in obvious ways for the case where some or all of the common items are internal to the tests that are being equated.

Designate the examinees who took both Tests X and W as Group 1; designate the examinees who took Tests Y and W as Group 2. Typically, every examinee falls in one of these two groups.

In practice, when there is a series of test forms A, B, ..., X, Y, Z, ..., say, the Group 1 data on Test X are processed as soon as they become available in order to equate Test X to the preceding form. When the Group 2 data become available at some later date, it is often considered uneconomical to rerun the Group 1 data, so Group 2 is run by itself. This case, where item parameters for Groups 1 and 2 are estimated separately, is the case to be considered here. The simplifying assumption that is used below to approximate the sampling variances of the estimated item parameters is not available in the alternative case where Groups 1 and 2 are pooled and all parameters estimated simultaneously.

New Equating Formulas

When parameters are estimated separately for Groups 1 and 2, the item parameters and θ in Equation 2 have a different origin and scale from the item parameters and θ in Equation 1. It is thus no longer possible simply to eliminate θ from Equations 1 and 2 to obtain the relation of η and ξ. The customary procedure in this situation is to use the anchor test to transform the Group 2 item parameters onto the scale of the Group 1 item parameters. This procedure adds to the sampling variance of the transformed item parameters and greatly complicates any determination of the sampling variance of the subsequent equating. The procedures and formulas given below avoid this problem, since they avoid any transformation of item parameters.

Equations 1 and 2 remain unchanged except that additional subscripts (explained below) are used. In particular, the symbols θ_1 and θ_2 must be distinguished because Groups 1 and 2 use different ability scales:

$$\xi = \sum_g P_{g1}(\theta_1) \quad , \qquad [3]$$

$$\eta = \sum_g P_{g4}(\theta_2) \quad . \qquad [4]$$

The item response functions here are written P_{gp} where $p = 1, 2, 3, 4$ refers to Test X, Group 1; Test W, Group 1; Test W, Group 2; and Test Y, Group 2, respectively, and $g = 1, 2, \ldots, n_p$, where n_p is the number of items in the appropriate test.

Now, similar equations will be written for the expected number-correct score ω on anchor Test W:

$$\omega = \sum_g P_{g2}(\theta_1) \quad , \qquad [5]$$

$$\omega = \sum_g P_{g3}(\theta_2) \quad . \qquad [6]$$

The desired equating relation between η and ξ can be obtained by eliminating θ_1, θ_2, and ω from these four equations.

Computer programs are available for equating η to ξ by eliminating θ from Equations 1 and 2. These same programs can be used to equate ω and ξ in one step, using Equations 3 and 5, then to equate η to ω in a second step using Equations 6 and 4. This produces an equating of η to ξ for the presently relevant situation where Group 1 and Group 2 parameters are not on the same scale.

An estimated equating is obtained from Equations 3 through 6 after replacing the true item parameters by their maximum likelihood estimates. Using carets to denote this change, the following equations are obtained:

$$\xi = \sum_g \hat{P}_{g1}(\theta_1) \quad , \qquad [7]$$

$$\hat{\omega} = \sum_g \hat{P}_{g2}(\theta_1) \quad , \qquad [8]$$

$$\hat{\omega} = \sum_g \hat{P}_{g3}(\theta_2) \quad , \qquad [9]$$

$$\hat{\eta} = \sum_g \hat{P}_{g4}(\theta_2) \quad . \qquad [10]$$

These equations show that $\hat{\eta}$ is a function of all the estimated item parameters together with the specified value of ξ.

Derivatives

For item g, write t_{1gp}, t_{2gp}, and t_{3gp} to denote the three parameters commonly used in IRT, instead of writing a_g, b_g, and c_g, respectively. Certain derivatives, obtained from Equations 4 through 6, will be needed for $r = 1, 2, 3$:

$$\frac{\partial \eta}{\partial t_{rg4}} = P^{(r)}_{g4}(\theta_2) \quad , \qquad [11]$$

$$\frac{\partial \omega}{\partial t_{rg3}} = P_{g3}^{(r)}(\theta_2) \quad , \qquad [12]$$

$$\frac{\partial \omega}{\partial t_{rg2}} = P_{g2}^{(r)}(\theta_1) \quad , \qquad [13]$$

where $P_{gp}^{(r)}$ denotes the derivative of P_{gp} with respect to t_{rgp}. Similarly,

$$\frac{\partial \eta}{\partial \theta_2} = \sum_g P'_{g4}(\theta_2) \quad , \qquad [14]$$

$$\frac{\partial \omega}{\partial \theta_1} = \sum_g P'_{g2}(\theta_1) \quad , \qquad [15]$$

where P' denotes a derivative with respect to θ. Using the formula for the derivative of an implicit function, it is found from Equations 3 and 6 for $r = 1, 2, 3$ that

$$\frac{\partial \theta_2}{\partial t_{rg3}} = - \frac{P_{g3}^{(r)}(\theta_2)}{\sum_g P'_{g3}(\theta_2)} \quad , \qquad [16]$$

$$\frac{\partial \theta_1}{\partial t_{rg1}} = - \frac{P_{g1}^{(r)}(\theta_1)}{\sum_g P'_{g1}(\theta_1)} \quad , \qquad [17]$$

$$\frac{\partial \theta_2}{\partial \omega} = \frac{1}{\sum_g P'_{g3}(\theta_2)} \quad . \qquad [18]$$

Using the chain rule for derivatives, it is found from the above formulas that

$$\frac{\partial \eta}{\partial t_{rg3}} = \frac{\partial \eta}{\partial \theta_2} \frac{\partial \theta_2}{\partial t_{rg3}} = - P_{g3}^{(r)}(\theta_2) \frac{\sum_g P'_{g4}(\theta_2)}{\sum_g P'_{g3}(\theta_2)} \quad , \qquad [19]$$

$$\frac{\partial \eta}{\partial t_{rg2}} = \frac{\partial \eta}{\partial \theta_2} \frac{\partial \theta_2}{\partial \omega} \frac{\partial \omega}{\partial t_{rg2}} = P_{g2}^{(r)}(\theta_1) \frac{\sum_g P'_{g4}(\theta_2)}{\sum_g P'_{g3}(\theta_2)} \quad , \qquad [20]$$

$$\frac{\partial \eta}{\partial t_{rg1}} = \frac{\partial \eta}{\partial \theta_2} \frac{\partial \theta_2}{\partial \omega} \frac{\partial \omega}{\partial \theta_1} \frac{\partial \theta_1}{\partial t_{rg1}} = - P_{g1}^{(r)}(\theta_1) \frac{\sum_g P'_{g2}(\theta_1)}{\sum_g P'_{g1}(\theta_1)} \frac{\sum_g P'_{g4}(\theta_2)}{\sum_g P'_{g3}(\theta_2)} \quad . \qquad [21]$$

Given ξ, it is now possible to express $\hat{\eta}$ as a series in powers of $\hat{t}_{rgp} - t_{rgp}$ ($r = 1, 2, 3$; $g = 1, 2, \ldots, n_p$; $p = 1, 2, 3, 4$):

$$\hat{\eta} = \eta + \sum_{p}\sum_{g}\sum_{r} (\hat{t}_{rgp} - t_{rgp})\eta'_{rgp} + \frac{1}{2}\sum_{p}\sum_{q}\sum_{g}\sum_{h}\sum_{r}\sum_{s} (\hat{t}_{rgp} - t_{rgp})(\hat{t}_{shq} - t_{shq})\eta''_{rgpshq} + \ldots \quad [22]$$

where η'_{rgp} is written instead of $\partial\eta/\partial t_{rgp}$ and η''_{rgpshq} instead of $\partial^2\eta/\partial t_{rgp}\partial t_{shq}$.

Sampling Variance

Transposing, squaring, and taking expectations, the following is found from Equation 22 for fixed ξ,

$$\text{Var}\ \hat{\eta} = \mathcal{E}(\hat{\eta} - \eta)^2 = \sum_{p}\sum_{q}\sum_{g}\sum_{h}\sum_{r}\sum_{s} \eta'_{rgp}\eta'_{shq}\ \text{Cov}(\hat{t}_{rgp}, \hat{t}_{shq}) + \ldots \quad [23]$$

When item parameters and abilities are both estimated simultaneously by maximum likelihood, it is not practical to use here the usual sampling covariance formulas for all estimators simultaneously. As a rough approximation, it is customary (Lord, 1980, Section 12.3) to use instead the (simpler) formulas for the case where the ability parameters are known. This rough approximation will be used here to find $\text{Cov}(\hat{t}_{rgp}, \hat{t}_{shq})$. Because of this approximation, the result will be an underestimate of the true sampling variance of equating.

In this case, all covariances involving two different items are exactly zero, as are all covariances involving a single item administered to two different groups of examinees. All nonzero variances and covariances are inversely proportional to N, the number of examinees. Thus,

$$\text{Var}\ \hat{\eta} = \sum_{p}\sum_{g}\left[\sum_{r=1}^{3}\sum_{s=1}^{3}\{\eta'_{rgp}\eta'_{sgp}\ \text{Cov}(\hat{t}_{rgp},\hat{t}_{sgp})\}\right.$$

$$\left. + \sum_{}^{3}\sum_{}^{3}\sum_{}^{3}\{\ \} + \sum_{}^{3}\sum_{}^{3}\sum_{}^{3}\sum_{}^{3}\{\ \} + \ldots\right] . \quad [24]$$

Some higher order terms are indicated here in order to make clear that the number of terms under summation signs does not increase too rapidly. The triple summation represents three times as many terms as the double summation, but each term in the triple summation is divided by $N^{3/2}$, whereas each term in the double summation is only divided by N. When N is several thousand, it is reasonable to expect that the higher order terms can be neglected, as is customary with asymptotic variances.

The final asymptotic formula is thus

$$\text{Var}\ \hat{\eta} \doteq \sum_{p=1}^{4}\sum_{g=1}^{n_p}\sum_{r=1}^{3}\sum_{s=1}^{3} \eta'_{rgp}\eta'_{sgp}\ \text{Cov}(\hat{t}_{rgp},\hat{t}_{sgp}) . \quad [25]$$

The η' values required here are computed from Equations 11, and 19 through 21. The covariances are obtained by the usual formulas for covariances of maximum likelihood estimators with fixed item parameters (Lord, 1980, p. 191).

Practical Application

Without data, it is difficult to make inferences about the magnitude of the sampling errors in IRT equating. Will they be larger or smaller than the sampling errors in conventional linear equating? In conventional equipercentile equating? Do sampling errors become large or small at extreme score levels?

Equation 25 was applied to an equating of the Verbal score on the 90-item Form VSA4 of the Scholastic Aptitude Test (SAT; December 1973 administration) to the 85-item Form XSA2 Verbal score (April 1975 administration). All examinees took an SAT and also a 40-item anchor test. Petersen, Cook, and Stocking (1980) made separate LOGIST runs on the 130 items in the 1973 administration for a sample of 2,665 examinees, and on the 125 items in the 1975 administration for a sample of 2,686 examinees; their item parameter estimates were used here.

SAT scaled scores are a linear transformation of the formula scores (corrects minus one-quarter incorrects). The results here are for the hypothetical case where all examinees answer all items. In this special case, formula scores are a linear transformation of number-correct scores; likewise, so are scaled scores. Since a known linear transformation $A\xi + B$ of number-correct scores ξ simply multiplies the standard error of $\hat{\eta}$ by the constant A, it is not difficult to obtain scaled-score standard errors from Equation 25.[1]

Table 1
A Comparison of Linear and IRT Equatings and of Their Standard Errors

Selected Formula Scores* XSA2	Linear Model		IRT Model	
	Equivalent Scaled Score	Standard Error	Equivalent Scaled Score	Standard Error
84	780	4.6	813.8	2.3
79.74	750	4.2	778.0	4.5
72.70	700	3.6	717.6	4.4
65.65	650	3.1	658.8	3.6
58.61	600	2.5	602.4	2.8
51.57	550	2.1	548.0	2.2
44.52	500	1.7	495.4	2.0
37.48	450	1.5	445.7	2.1
30.43	400	1.6	399.3	2.3
23.39	350	1.8	355.6	2.8
16.35	300	2.3	313.3	3.6
9.30	250	2.8	270.2	4.7
2.26	200	3.3	223.0	7.0
-5	150	3.9	163.5	15.6

*Although formula score is actually a discrete variable, it is for convenience treated here as continuous.

For each of certain specified formula scores on Form XSA2, Table 1 shows (1) the equivalent scaled score found by the conventional linear procedure usually used for the SAT (Design IVA;

[1]A computer program to do this was written and run by Marilyn Wingersky.

Angoff, 1971); (2) the standard error of these equated (scaled) scores as found by the computer program AUTEST (Lord, 1975), assuming the validity of the linear model; (3) the equivalent scaled score found by the IRT method of this article; and (4) the corresponding scaled-score standard error calculated from Equation 25. The standard errors in Table 1 are best understood in comparison with the standard deviation of scaled scores, which is 106 for XSA2, and in comparison with the classical test theory standard error of measurement (due to imperfect reliability), which is 31. Clearly, the standard error of equating is small compared to the standard error of measurement.

Judging by the IRT standard errors, the equating is definitely nonlinear, at least outside the score range from 350 to 650. The IRT standard errors show a continued sharp increase as the minimum possible true formula score of −5.5 is approached. At the other end of the score scale, the IRT standard error increases up to a scaled score of 760 and decreases thereafter. The reason for the decrease at the upper end is that for a perfect score, the standard error of this kind of IRT equating is zero. Except at the upper end, the IRT standard error is larger than that of the linear model.

The results of Table 1 are displayed in Figures 1 and 2. The straight line in Figure 1 shows the linear equating of true formula score on XSA2 to true scaled score on VSA4. The dashed lines are drawn two standard errors above and below the straight line.

Figure 1
Linear Equating of True Formula Score on XSA2 to True Scaled Score on VSA4
(Dashed Lines are Two Scaled-Score Standard Errors Above and Below Equating Line)

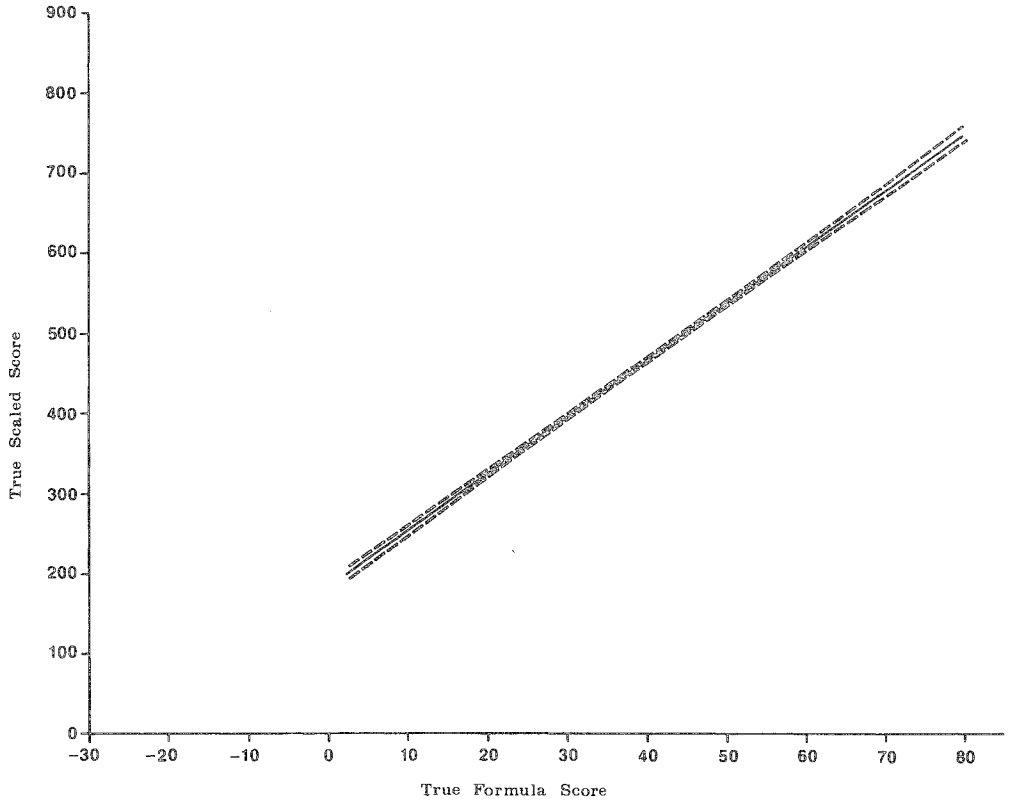

Figure 2 similarly displays the curvilinear IRT equating of XSA2 to VSA4 and its standard error. The straight line extension of the lower end of the equating (middle) line in Figure 2 was obtained by the method described in Lord (1980, pp. 210-211). It is shown in the figure for completeness; but no standard error is shown, since there is no good theoretical basis for such an extension.

Figure 2
IRT Equating of XSA2 Formula Score to VSA4 Scaled Score,
with Two-Standard-Error Bounds

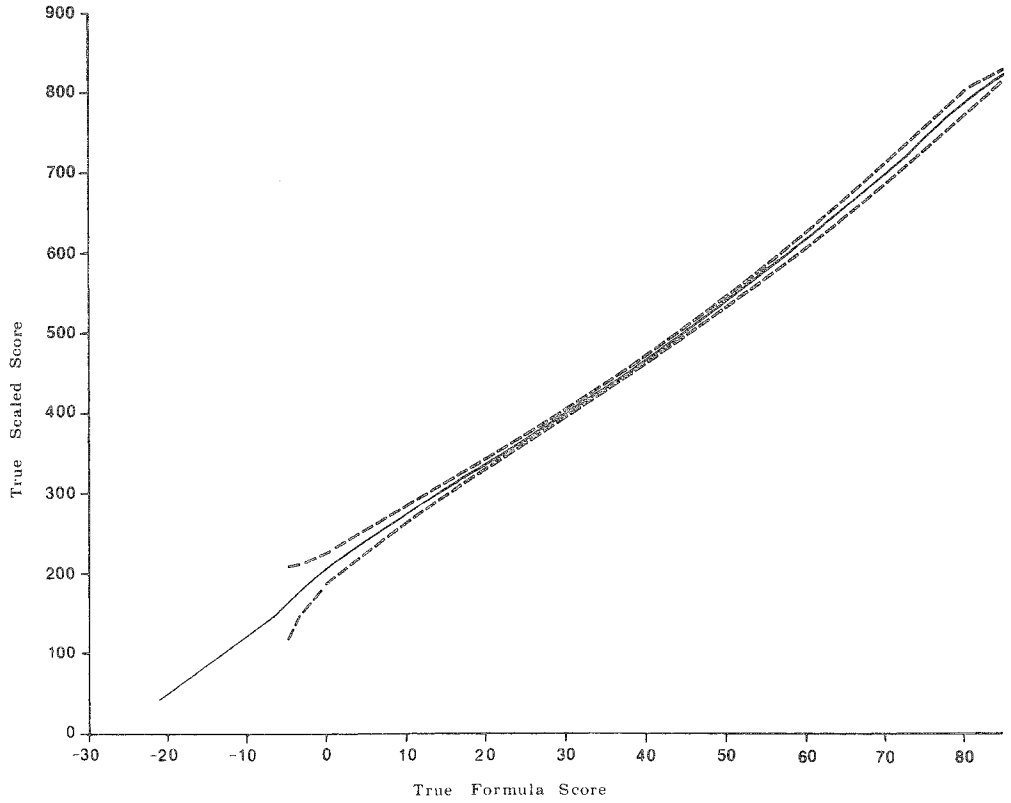

Table 2 compares present IRT equating with a conventional equipercentile equating of XSA2 to VSA4 via the anchor test. In conventional equating, an XSA2 score and a VSA4 score each equipercentilely equivalent to a given anchor test score are considered to be equivalent to each other. The standard error of the resulting equipercentile equating of XSA2 to VSA4 is given by $(SE^2_{XSA2} - SE^2_{VSA4})^{1/2}$, where the SE are standard errors of separate equipercentile equatings of each test to the anchor test. Formulas for SE_{XSA2} and SE_{VSA4} are given in Lord (in press).

Since SE_{XSA2} and SE_{VSA4} are estimated from unsmoothed data, the equipercentile standard errors in Table 2 fluctuate somewhat. Nevertheless, it is apparent that the equipercentile method has a much larger standard error above a scaled score of 450. For these data, the IRT method shows a larger standard error than the equipercentile method only when the formula score is negative.

Table 2
A Comparison of Equipercentile and IRT Equating and of Their Standard Scores

XSA2 Formula Score	Equipercentile Method		IRT Model	
	Equivalent Scaled Score	Standard Error	Equivalent Scaled Score	Standard Error
78.1	774	13.47	764	4.68
70.6	722	15.85	700	4.18
64.75	652	10.32	651	3.44
58.9	602	4.97	605	2.78
52.9	558	4.12	558	2.32
47.25	514	3.47	515	2.09
40.1	466	3.44	464	2.05
32.4	417	2.93	412	2.24
25.75	364	3.37	370	2.63
16.1	314	4.07	312	3.62
7.6	242	5.70	259	5.08
-3.75	195	7.85	175	12.49

The standard error of equipercentile equating could be reduced by smoothing the frequency distribution of raw scores before equating. Smoothing is undoubtedly desirable as a practical expedient; however, the choice of a smoothing formula is somewhat arbitrary and the smoothing is likely to pre-

Table 3
IRT Equatings and Their Scaled-Score Standard Errors, a Comparison of Results Using 20- and 40-Item Anchor Tests

XSA2 Formula Score	Length of Anchor Test			
	20 Items		40 Items	
	Scaled Score	Standard Error	Scaled Score	Standard Error
80	787	5.9	780	4.5
70	698	5.3	695	4.1
60	615	3.9	613	2.9
50	540	3.0	536	2.2
40	467	2.7	463	2.0
30	399	3.0	397	2.4
20	336	3.9	335	3.2
10	274	5.4	275	4.6
0	206	9.9	206	8.4

vent convergence of the estimated equating to its true value in large samples. Formulas for the standard errors of smoothed equipercentile equating are not presently available.

In order to determine the effect of using a shorter anchor test, every other item in the anchor test was discarded and the data reanalyzed on the basis of the remaining 20-item anchor test. The effect on the standard errors of IRT equating is shown in Table 3. The two equatings agree fairly well. At the point where the equating standard errors are a minimum, halving the length of the anchor test increases the standard error by a factor of about $\sqrt{2}$. At the other score points, the effect is less. Given standard errors like those in Table 2, it will now be possible to make a reasonable judgment as to the length necessary for an anchor test.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington DC: American Council on Education, 1971.

Cowell, W. R. *ICC pre-equating in the TOEFL testing program*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Lord, F. M. Automated hypothesis tests and standard errors for nonstandard problems. *The American Statistician*, 1975, *29*, 56-59.

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. *Journal of Educational Statistics*, in press.

Petersen, N. S., Cook, L. L., & Stocking, M. S. *Scale drift: A comparative study of IRT versus linear equating methods*. Paper presented at the Fourth International Symposium on Educational Testing, Antwerp, Belgium, June 1980.

Petersen, N. S., Cook, L. L., & Stocking, M. L. *IRT versus conventional equating methods: A comparative study of scale stability*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Yen, W. M. *Use of three-parameter item response theory in the development of CTBS, Form U, and TCS*. Paper presented at the annual meeting of the American Educational Research Association, New York, March 1982.

Acknowledgments

This work was supported in part by Contract N00014-80-C-0402, Project NR 150-453, between the Office of Naval Research and Educational Testing Service.

Author's Address

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