A Study of Pre-Equating Based on Item Response Theory

Isaac I. Bejar and Marilyn S. Wingersky
Educational Testing Service

The study reports a feasibility study using item response theory (IRT) as a means of equating the Test of Standard Written English (TSWE). The study focused on the possibility of pre-equating, that is, deriving the equating transformation prior to the final administration of the test. The three-parameter logistic model was postulated as the response model and its fit was assessed at the item, subscore, and total score level. Minor problems were found at each of these levels; but, on the whole, the three-parameter model was found to portray the data well. The adequacy of the equating provided by IRT procedures was investigated in two TSWE forms. It was concluded that pre-equating does not appear to present problems beyond those inherent to IRT-equating.

Equating, in general, refers to the derivation of transformations that map scores on different forms of a test onto a scale in such a way that after transformation the scores on the various forms are comparable. The equating methodology that has been commonly used (see Angoff, 1971) requires that the form being equated first be administered to testees. Since, in large-scale testing programs, scores are not due back to testees for 4 to 6 weeks, it would seem that there is ample time to derive the equating transformation. In practice, the bulk of the time is consumed by various data-processing steps. As a result, the equating transformation must be produced in a rather short period of time. Even when no difficulties arise, the psychometrician is under considerable pressure.

From this pragmatic point of view, one of the most promising applications of item response theory (IRT) is pre-equating (see Lord, 1980, chap. 13). As implied by the name, pre-equating refers to the derivation of the transformation prior to the administration of the form to be equated. This requires that IRT item statistics be available on a common metric for *all* the items that appear in the final form. The feasibility of implementing pre-equating for the Test of Standard Written English (TSWE) using the three-parameter logistic model (Birnbaum, 1968) is the focus of the present study. Although some theoreticians (e.g., Samejima, 1973) have criticized the model on a number of points, at present, it seems to be the most practical one. Although IRT would seem essential for item level pre-equating, in some circumstances "section-pre-equating" may be more appropriate (see Holland & Wightman, in press).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 3, Summer 1982, pp. 309–325

0146-6216/82/030309-17$1.85

Overview of the Study

Whether pre-equating works or not depends on two factors. One is the fit of the three-parameter model to TSWE data. Since there is no general procedure for ascertaining fit, several procedures were used in the hope that collectively they can be more revealing. The second factor that may prevent successful pre-equating is lack of "situational" invariance in the item parameter estimates. In practice, pre-equating requires that the final form be assembled from items coming from various pretest forms. This raises the possibility of a context effect on item parameters, which (as shown by Yen, 1980), can be substantial. The adequacy of pre-equating was judged in this study on two forms in which these conditions could be simulated, using as a criterion scores equated by means of non-IRT procedures.

Method

Data and Calibration Procedures

Description of the TSWE. The TSWE is a 30-minute multiple-choice test administered together with the Scholastic Aptitude Test (SAT). Its purpose is to help colleges place students in appropriate English Composition courses. It is not recommended as an admissions instrument. The test consists of 50 items; Items 1 to 25 and 41 to 50 are called usage items, while Items 26 to 40 are called sentence correction items. A more complete description of the test can be found in Bejar and Wingersky (1981) and in Breland (1976).

Item calibration procedures. Because of the expense involved in item calibration, the adequacy of pre-equating was investigated for only two TSWE forms: E7 and E8. As will be seen, however, to obtain item statistics on even two forms is not straightforward. The calibration of a large set of items administered to different samples involves (1) obtaining item parameter estimates on the arbitrary metric defined by each calibration sample and (2) placing all items calibrated on different samples on the same metric.

Parameter estimation. All item parameter estimates used in this report were obtained using the program LOGIST (Wood, Wingersky, & Lord, 1976); the three-parameter logistic model (Birnbaum, 1968) was the assumed response function. The function of the LOGIST program is to estimate, for each item, the three item parameters: discrimination (a), difficulty (b), and a pseudo-guessing parameter (c). Unless otherwise indicated, the following constraints were imposed on the estimation: a was restricted between .01 and 1.25; c was held fixed to .15 until Stage 2 of Step 2. Thereafter, a, b, and c are estimated except that c's were held fixed at a constant, $\bar{c}$, estimated by the program for those items with $b - 2/a < -2.0$ at the end of Stage 3 of Step 2 (Wood et al., 1976). The c's for all other items were restricted to a range of .0 to .5.

Parameter estimates were based on spaced samples of 3,000 students chosen for each data set. However, a few subjects were excluded from each sample for a variety of reasons.

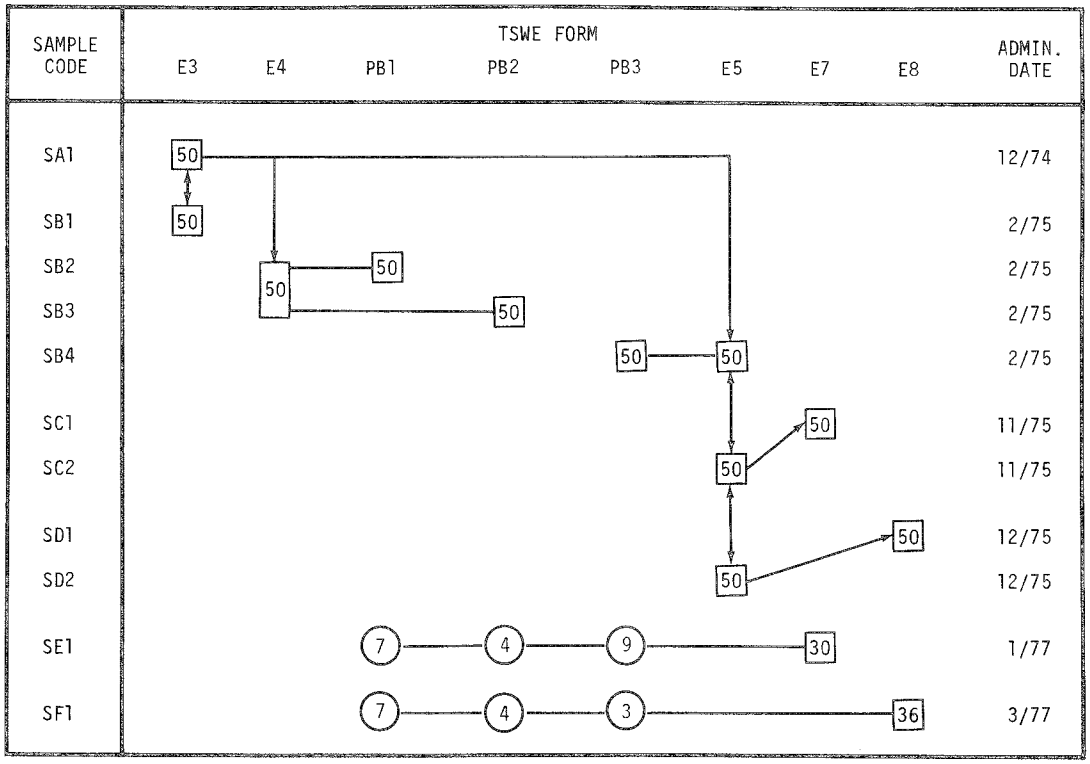
Putting estimates on the same metric. Two procedures were used to place estimates on the same metric. One procedure sets the scale of the items being calibrated by fixing the b estimates of the items previously calibrated and re-estimating the a and c. Obviously, this requires that the previously calibrated items already be on the desired metric and that they be administered together with the items being calibrated.

The above procedure can be used when previously calibrated items are administered together with uncalibrated items. The second procedure puts item statistics on the desired scale by applying a linear transformation to the item statistics. The procedure requires that the uncalibrated items, or new form, be administered by themselves to a random sample of population X and that the previously

calibrated items, or old form, also be administered to a random sample of population X. The items are then calibrated separately for the new form and for the old form. The new form is put onto the scale of the old form in the new administration by setting the means and standard deviations of the abilities to be equal. There are now two separate estimates of the b parameters for the old form: one from the new administration and one from a previous administration. If the model holds, these estimates are linearly related. A variety of procedures can now be used to derive the linear relationship to transform the b's for the new administration, old form, onto the scale of the previous administration. For example, the mean and standard deviations of the two sets of b estimates can be equated. However, in this report a robust procedure was used. This procedure, adapted by Lord, is explained in the Appendix. Once the transformation is derived, it is applied to the a and b parameter estimates of the new form.

Figure 1 shows the calibration procedure. The metric was defined with respect to SA1-E3, i.e., Sample A1 and Form E3. A double-headed arrow indicates the derivation of the transformation described in Appendix A. A single-headed arrow indicates the application of that transformation. A connecting line is used to indicate that items were administered to the same sample. The number within a square indicates the number of items for which a, b, and c were estimated. The number within a circle indicates the number of items for which b was fixed and a and c estimated. Forms PB1, PB2, and PB3 were pre-test forms. E7 contained 20 items from three pretest forms, while E8 contained 14. For example, for the SE1-E7 data set the b's of 20 items were fixed to values obtained with combined Data Sets SB2-PB1 and SB3-PB1, and Data Set SB4-PB3.

Figure 1
Data Used in the Calibration



Fit of the Three-Parameter Model to the TSWE Data

The following procedures were used to examine the fit of the three-parameter logistic model to TSWE data:

1. Examining the estimated regression of item score on ability.
2. Contrasting observed and expected distribution of number-correct scores.
3. Examining the factorial structure of the TSWE.

Evaluation of fit at the item level. An intuitively appealing way to examine fit is by comparing the "observed" regression of item score on ability against the estimated regression of item score on ability predicted by the model, i.e., the item characteristic curve (ICC). The comparison permits a visual assessment of how well the estimated parameters portray the response data for a given item. For the present application plots were constructed for each item, as follows. The estimated ICC is given by

$$P_i(\theta) = \hat{c}_i + (1 - \hat{c}_i)\left\{1 + \exp\left[-1.7\hat{a}_i(\theta - \hat{b}_i)\right]\right\}^{-1} \quad [1]$$

where $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ are the estimated item parameters for Item i and $P_i(\theta)$ is the probability of answering the item correctly for someone of ability θ.

The "observed" regression of item score on ability was computed by dividing θ into intervals of .4 and grouping students into those intervals based on their $\hat{\theta}$. (This is an approximation, since $\hat{\theta}$ contains errors.) Within the k^{th} interval the probability of a correct response was computed as

$$P^*_{ik} = \left[R_k + O_k/A\right] / N_k \quad [2]$$

where

R_k is the number of testees who answered the item correctly in the k^{th} interval,
O_k is the number of students who omitted the item in the k^{th} interval,
A is the number of alternatives in the item, and
N_k is the total number of testees in the k^{th} interval.

For purposes of this study, it is of most interest to examine the plots for items where the b parameter had been fixed to their pretest value. It was found that with few exceptions, fixing the b's did not erode the fit. Figure 2 shows the plots for the two poorest cases. The squares constitute the "observed" item-on-ability regression; the solid curve is the estimated ICC. The size of the square is proportional to the number of testees in that interval of θ. The asterisk next to the b value indicates that the b parameter was fixed to that value.

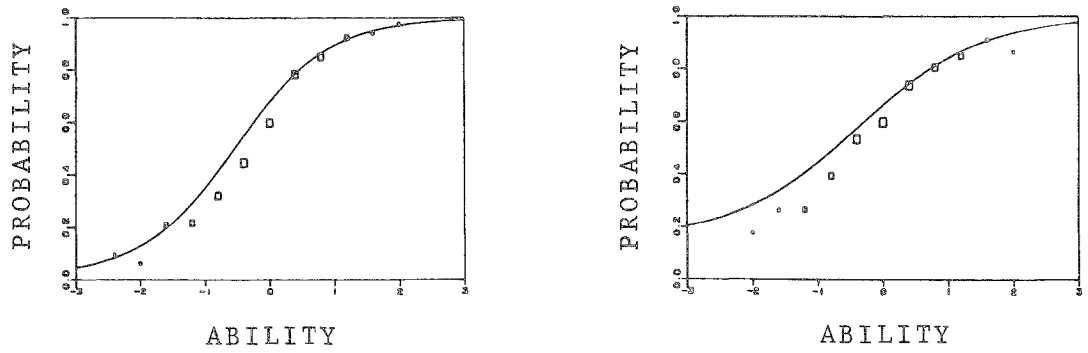
Evaluation of fit at the total score level. A logical extension of the previous procedure is to consider how well the model predicts the distribution of number-correct scores in a given sample. Since the three-parameter model has no way of predicting omits for a given individual, the analysis is based on number-correct score rather than formula scores.[1] The rationale of this procedure is to compare a prediction of the model (in this case, the frequency distribution of number-correct scores) against the empirical results (in this case, the observed distribution of number-correct scores). Although a number of indices could be used to quantify the discrepancies between the predicted and observed distributions, none were used; therefore, the assessment of fit was judgmental.

The observed frequency distribution of number-correct scores was obtained by simply tabulating the number of testees at each number-correct score level. The predicted frequency distribution was obtained by a complex algorithm; however, its conceptual equivalent is easily understood as follows:

1. For each testee determine n^*, the number of items reached.

[1]The TSWE is scored operationally using scores corrected for guessing. Such scores are referred to as formula scores.

Figure 2
Observed Regressions of Item Score on Ability and Estimated ICC for the Poorest Fitting Items from Forms E7 and E8



TSWE.SE1-E7 Item 23
A=0.8133 B=-0.5267*
C=.0151 R-BIS=0.6019

TSWE.SF1-E8 Item 34
A=0.6195 B=-0.3907*
C=.1530* R-BIS=0.4324

2. Compute the $P_i(\theta_a)$ and $Q_i(\theta_a)$ where $P_i(\theta_a)$ is the probability of answering the i^{th} item correctly, as given by the three-parameter logistic model, for a given θ_a: $Q_i(\theta_a) = 1 - P_i(\theta_a)$.
3. Generate all possible 2^{n*} response vectors such that $u_i = 1$ indicates a correct response and $u_i = 0$ indicates an incorrect response.
4. For each vector substitute $P_i(\theta_a)$ if $u_i = 1$ and $Q_i(\theta_a)$ if $u_i = 0$; multiply the probabilities to obtain the probability of the response vector. That is, compute

$$\prod_{i=1}^{n^*} \hat{P}_i(\theta_a)^{u_i} \hat{Q}_i(\theta_a)^{(1-u_i)} \qquad [3]$$

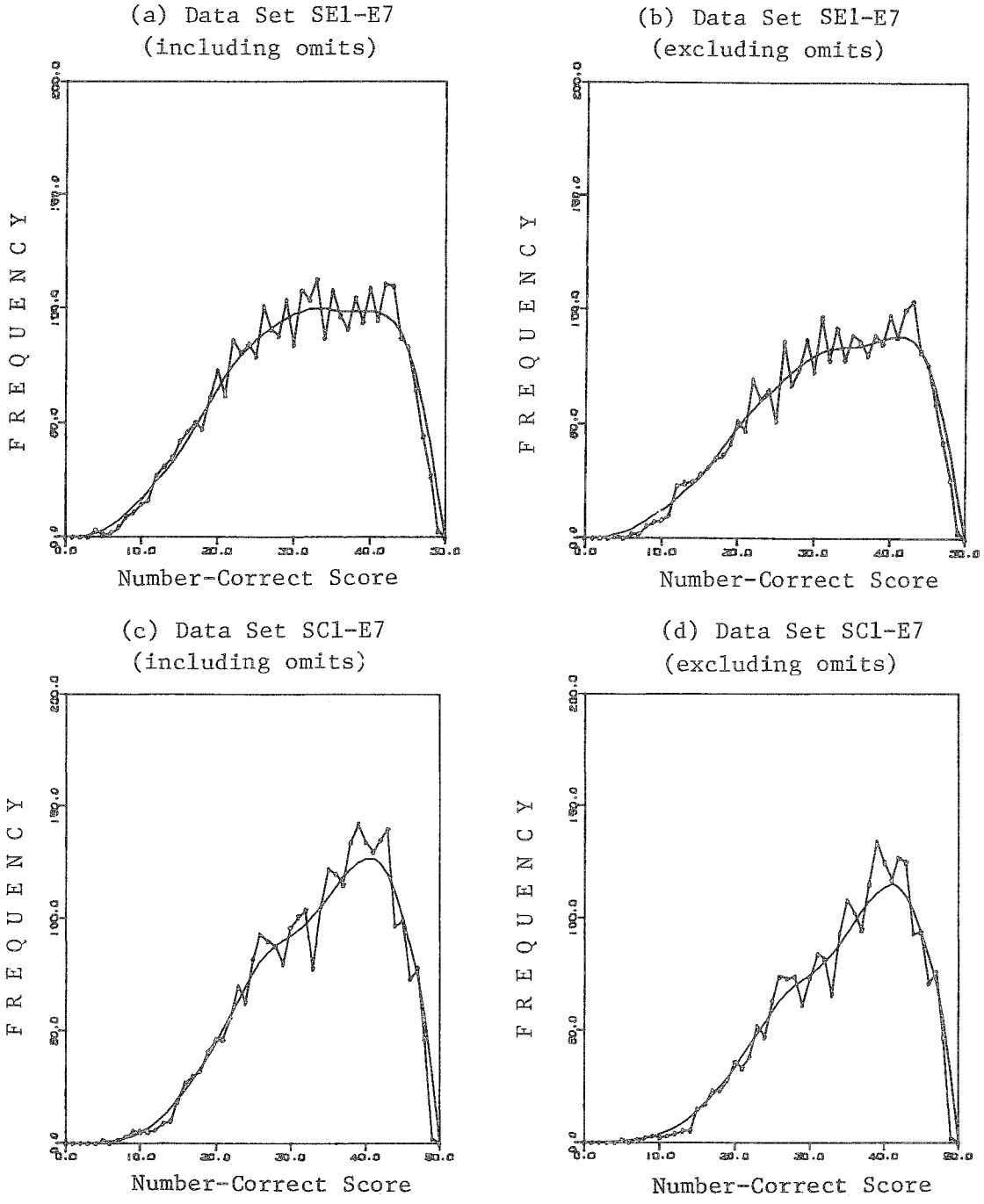
5. Group response vectors with the same number of 1's. There are $n^* + 1$ such groups corresponding to number-correct scores of 0, 1, 2, . . ., n^*.
6. Sum the probabilities of each response vector within a group. The sum of these probabilities is the expected frequency of this number-correct score. When done for each group, this gives the expected distribution of number-correct score for one testee.
7. Repeat the above steps for each testee and sum the distribution over examinees for each number-correct score.
8. Divide by N, the number of testees, which yields the expected distribution of number-correct scores for the entire sample.

Note that this procedure assumes local independence, since the product of probabilities is taken in the fourth step, which is the reason why the comparison against the observed distribution of number-correct scores may be viewed as a test of fit.

The procedure was applied once excluding testees with omits and once including those students. The results for Form E7 are shown in Figure 3. As can be seen, the discrepancies between observed and expected distributions do not appear any larger when some of the b's have been fixed, as in Data Set SE1-E7.

Factorial structure of TSWE data. The third method of assessing fit involves factor analysis. Attempts to examine fit through factor analysis (e.g., Indow & Samejima, 1966) have used inter-item

Figure 3
Plot of the Observed and Expected Distribution
of Number-Correct Scores for Form E7



correlation matrices. By contrast, the present use of factor analysis involves correlations among subscores. Since the TSWE contains two item types, a reasonable hypothesis is that response to each item type requires somewhat different processes. That is, the two item types do not measure the same construct.

Two formula scores were computed for each item type by totaling across odd and even items separately. To insure that the odd and the even scores were based on the same number of items, Item 25 was excluded from the odd items for the usage items and Item 40 was excluded from the even items for the sentence correction items.

Correlation and covariance matrices were computed based on the four scores for the following data sets: SA1-E3, SB2-E4, SB4-E5, SC1-E7, and SD1-E8. A two-factor model was fitted to each of these correlation matrices and the parameters were estimated by the maximum likelihood procedure using the COFAMM Program (Sörbom & Jöreskog, 1976). The model tests the hypothesis that the correlation matrix, Σ, is described as follows:

$$\Sigma = \begin{bmatrix} x_1 & 0 \\ x_2 & 0 \\ 0 & x_3 \\ 0 & x_4 \end{bmatrix} \begin{bmatrix} 1 & x_5 \\ x_5 & 1 \end{bmatrix} \begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_3 & x_4 \end{bmatrix} + \begin{bmatrix} x_6 & 0 & 0 & 0 \\ 0 & x_7 & 0 & 0 \\ 0 & 0 & x_8 & 0 \\ 0 & 0 & 0 & x_9 \end{bmatrix}$$

the x's are parameters to be estimated. The 0's indicate the corresponding parameter is fixed to zero. x_1 through x_4 represent the factor loadings; x_5 is the correlation among the two factors, each defined by an item type; and x_6 through x_9 represent the unique variance of each subscore. This model is discussed by Jöreskog (1978), who notes it is a restatement of an earlier model by Lord (1956).

For purposes of this study, the model can be used to test the presence of an item type effect by estimating the model with x_5 set to 1.0, that is, by hypothesizing the correlation among true scores to be perfect. This was done for the five data sets mentioned earlier and, in every case, the hypothesis $x_5 = 1$ was rejected with $p < .0001$. The model was then estimated allowing x_5 to be estimated. The results are shown in Table 1. For all forms but E8 the model fit ($p > .05$). Even when the p values are not accurate (since the data does not have a multivariate normal distribution, as assumed by COFAMM), the magnitude of the chi-square statistic suggests that for E8 the two-factor models did not account as well for all the correlations.

The estimated correlation between the true scores for the two item types is also shown in Table 1. The asymptotic standard error of the estimated correlation is shown in parentheses. As can be seen, the correlations are below .90 except for E8.

This analysis suggests that the structure of the TSWE can be understood by postulating two item-type factors. Although the correlation between the two scores is very high, it does not approach 1.0 as would be expected if both scores measured a single construct. More concretely, removing the constraint that the two factors correlate 1.0 reduced the residuals to almost zero. To illustrate, the residual covariance matrices for SA1-E3 under the two models are shown in Table 2. The residuals under the hypothesis $x_5 = 1$ are shown above the diagonal. The corresponding residuals when x_5 is estimated are shown below the diagonal. (The first letter, *E* or *O*, stands for even or odd, respectively; US stands for usage, and SC stands for sentence correction.)

Table 1
Summary Results of Factor Analysis with Two Item Type Factors.

Form	Sample	χ^2	df	p	Correlation between True Scores[a]
W506	E3	.42	1	.52	.889 (.01)
X104	E4	2.77	1	.10	.884 (.01)
X106	E5	3.50	1	.06	.891 (.01)
X406	E7	.15	1	.70	.879 (.01)
X506	E8	18.04	1	.00	.915 (.01)

[a]The value in parenthesis is the asymptotic standard error of the correlation.

Assessment of Pre-equating

The criterion for judging adequacy of pre-equating in the present study is by comparison to conventional equating. Implicit in this choice of criterion is the assumption that conventional equating provides a reasonable criterion. While this is not generally true, the conventional equating was done by spiralling the old and new forms in random samples of the new population. Furthermore, the test specifications are observed very strictly and, as a result, there is minimal variation across forms of the test. All of this suggests conventional linear equating should work adequately with the TSWE. Nevertheless, three conventional equatings were used as criteria. One criterion, C1, was the operational linear equating, that is, the procedure used in the reporting of scores. For E7 and E8 the operational equatings used Scholastic Aptitude Test-Verbal section (SAT-V) and Mathematical section (SAT-M) as anchor tests. The second criterion used, C2, was similar to C1 except that only SAT-V was used as an anchor. Finally, the third criterion, C3, was equipercentile equating using SAT-V as an anchor test. (It would have been desirable to use SAT-V and SAT-M as anchors for equipercentile equating also, but the computer program did not allow it.) A description of linear and equipercentile equating can be found in Petersen, Cook, and Stocking (1981).

The comparison of pre-equating and operational equating results was limited to two forms, E7 and E8, since only on these forms was it possible to simulate pre-equating conditions. For operational use, that is, to actually report converted scores to testees, E7 had been equated to E5 in the 11/75 ad-

Table 2
Residual matrices for SA1-E3 under the hypothesis $x_5 = 1$, above the diagonal, and when x_5 is estimated, below the diagonal

	OUS	EUS	OSC	ESC
OUS	-	.012	-.016	-.014
EUS	.000	-	-.013	-.019
OSC	-.002	.002	-	.083
ESC	.002	-.002	.000	-

ministration; E8 had been also equated to E5 but in the 12/75 administration. In both cases, the "old" and "new" forms were spiralled, and scores on the SAT-V and SAT-M were used to adjust the TSWE scores before equating the TSWE means and standard deviations of the two forms. The results of conventional linear equating are two parameters, usually referred to as A and B, which are used for converting formula scores to the TSWE metric as follows:

$$S = A\,(F) + B \qquad [4]$$

where S is the converted score and FS is the formula score. Since the TSWE has 50 items, FS ranges from -12 to 50. If S is less than 20, it is set to 20. Also, if S becomes greater than 60, it is set to 60. The results of equipercentile equating is a table which converts scores in the old form to corresponding scores in the new form.

For methodological as well as practical reasons two levels of pre-equating were studied. The least demanding level consists of estimating IRT parameters for the new form (E7 or E8, in this case) when the items appear together as a form. Strictly speaking, this is not pre-equating but merely IRT-based true-score equating and will be referred to as IRT-equating. (If the IRT parameters had been estimated on a different population, it could be considered truly pre-equating). Nevertheless, precisely because it is a very undemanding form of pre-equating, the results from this comparison serve as a good benchmark to compare the results of pre-equating proper.

For the second level, pre-equating proper, the parameter sets E7P and E8P were used. For E7P and E8P the a's, b's, and c's of 20 and 14 items, respectively, were taken from parameter sets PB1, PB2, and PB3. For the remaining items the a's, b's, and c's were taken from the parameter set SE1-E7 and the parameter set SF1-E8. Within IRT-equating and pre-equating three old forms were used, namely, E3, E4, and E5, to put the new form E7 or E8 on the TSWE scale.

Equating Based on IRT

The procedure used to transform formula scores on the new form to scaled scores can be described in general as follows: For a given true score on the new form, find the corresponding θ. Next, find the true score on the "old" form associated with this θ. Finally, apply existing conversion parameters to put the equated true scores on the TSWE scale. Since the TSWE is scored using formula scores, the actual procedure is based on true formula scores. A step-by-step description follows.

1. For each integer formula score on the new form, FS_{new}, greater than $\ell = \sum_{i=1}^{50} c_{i\cdot new}$ and less than or equal to 50 (where $c_{i\cdot new}$ is the c estimated parameter for the i^{th} item), compute the associated true score number-correct scale as follows:

$$R_{new} = .80\,F_{new} + 50/5 \qquad [5]$$

   This is based on the fact that if an examinee attempts all items in the test, number-correct and formula score are linearly related with slope $m/(m-1)$, where m is the number of alternatives, and constant $n/(m-1)$, where n is the number of items in the test. A similar relationship holds for true formula scores and true scores as shown by Lord (1980, chap. 15).
2. The next step is to find the θ associated with a given R_{new}. This is done by solving for θ in the equation

$$R_{new} = \sum_{i=1}^{50} P_i^*(\theta) \qquad [6]$$

   where the $P_i^*(\theta)$ is computed using the $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, for the new form.
3. Having found the needed θ, compute the corresponding true score in the old form as follows:

$$R_{old} = \sum_{i=1}^{50} P_i(\theta) \quad [7]$$

Where now $P_i(\theta)$ is computed using the a_i, b_i, and c_i from the old form.

4. The true formula score corresponding to this true score is

$$F_{old} = R_{old}/.80 - 50/4 \quad [8]$$

5. Finally, F_{old} is converted by means of existing parameters A and B as follows:

$$S = A(F_{old}) + B \quad [9]$$

If F_{old} is less than ℓ, a somewhat different procedure is used (see Lord, 1980, chap. 13, Appendix C).

This procedure was applied with E3, E4, and E5 as old forms and new forms E7 and E8 (using parameter sets SC1-E7 and SD1-E8 for IRT-equating, and parameter sets E7P and E8P for pre-equating).

Table 3

Summary Conversion Table Comparing Conventional Equating, IRT-equating and Pre-equating for E7

	Criterion			IRT-equating Old form			Pre-equating Old form		
Raw Score	C1	C2	C3	E3	E4	E5	E3	E4	E5
50	60	60	60	60.	60.	60.	60.	60.	60.
45	58	58	59	58.	58.	59.	57.	57.	58.
40	53	53	53	54.	54.	54.	53.	52.	52.
35	48	48	48	49.	49.	49.	48.	48.	48.
30	43	43	43	45.	44.	44.	44.	44.	43.
25	38	38	38	40.	39.	39.	40.	39.	39.
20	33	33	33	35.	34.	34.	35.	34.	34.
15	28	28	28	30.	28.	28.	30.	29.	29.
10	22	22	22	24.	23.	23.	25.	24.	24.
5	20	20	20	20.	20.	20.	20.	20.	20.
0	20	20	20	20.	20.	20.	20.	20.	20.
-5	20	20	20	20.	20.	20.	20.	20.	20.
-10	20	20	20	20.	20.	20.	20.	20.	20.
Mean	43.84	43.83	43.69	45.25	44.55	44.78	44.66	44.06	44.07
S. D.	9.70	9.73	9.80	9.20	9.64	9.81	8.62	8.93	9.05

C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as an anchor; C3 is based on equi-percentile equating. The means and standard deviations are based on the formula score frequency distribution for the first national administration of E7.

Results

The magnitude of the discrepancies can be appreciated by examining the mean and standard deviation corresponding to the criterion equatings, IRT-equating, and pre-equating. The mean and standard deviations are based on the frequency distribution observed in the first national administra-

Table 4
Summary Conversion Table Comparing Conventional Equating, IRT-equating and Pre-equating for E8

	Criterion			IRT-equating Old form			Pre-equating Old form		
Raw Score	C1	C2	C3	E3	E4	E5	E3	E4	E5
50	60	60	60	60.	60.	60.	60.	60.	60.
45	59	59	60	58.	59.	60.	58.	58.	59.
40	53	53	54	53.	53.	53.	54.	53.	54.
35	47	48	48	49.	48.	48.	49.	49.	48.
30	42	42	42	43.	43.	42.	44.	43.	43.
25	36	36	35	38.	37.	37.	39.	38.	38.
20	30	30	29	33.	31.	32.	34	33	33.
15	24	24	24	27.	26.	26.	29.	27.	28.
10	20	20	20	22.	21.	21.	23.	22.	22.
5	20	20	20	20.	20.	20.	20.	20.	20.
0	20	20	20	20.	20.	20.	20.	20.	20.
-5	20	20	20	20.	20.	20.	20.	20.	20.
-10	20	20	20	20.	20.	20.	20.	20.	20.
Mean	42.10	42.14	42.04	43.59	42.84	42.94	44.30	43.60	43.59
S. D.	9.96	9.98	10.21	8.77	9.26	9.31	8.34	8.78	8.77

C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as an anchor; C3 is based on equi-percentile equating. The means and standard deviations are based on the formula score frequency distribution for the first national administration of E8.

tion of E7 (Table 3) and E8 (Table 4). In particular, four trends are more or less obvious. First, operational and IRT-based equatings are much more discrepant for E8 than for E7. Secondly, for the IRT-based conversions the mean is higher and the standard deviation smaller compared to the criterion equatings. Thirdly, the choice of an old form seems to affect the discrepancy of IRT-based and operational equating. More concretely, using E3 as an old form for either pre-equating or IRT-equating yields the most discrepant results. Finally, comparing the results for pre-equating and IRT-equating, for E7 pre-equating actually yields less discrepant results than IRT-equating but the opposite is true for E8.

A more detailed analysis of the results can be obtained from an index suggested by Marco, Petersen, and Stewart (1980), namely, the weighted mean squared error,

$$\Sigma_j f_j d_j^2/N = \Sigma f_j (d_j - \bar{d})^2/N + \bar{d}^2 \quad \text{or} \qquad [10]$$

(Total Error) = (Variance of Differences) + (Squared Bias)

where

$d_j = (t'_j - t_j)$;

t'_j is the criterion score (which in this case corresponds to the operational score) for raw score x_j;

t_j is the IRT-based converted score corresponding to the same raw score x_j;
$\overline{d} = f_j d_j / N$;
f_j is the frequency of x_j; and
$N = \Sigma f_j$.

Tables 5 and 6 show the computed indices for E7 and E8, respectively. An examination of the discrepancy indices largely corroborates the results noted earlier. Within a given equating procedure, there is variation due to the choice of old form. For IRT-equating E3 yields the most discrepant results in terms of the weighted mean squared differences, E4 the least discrepant results, and E5 is in between. This is true for both E7 and E8. For pre-equating E3 also yields the most discrepant results, but E5 yields the least discrepant results, with E4 in between. Again, this is true for both E7 and E8.

For E7, the linear equating using SAT-V and SAT-M as anchors (C1) yields the least discrepant results for both IRT-equating and pre-equating, followed by linear equating using only SAT-V as anchor (C2) and equipercentile equating (C3). For E8, however, using linear equating with only SAT-V as anchor (C2) yields the least discrepant results, followed by C1 and C3, in that order. That is, using equipercentile equating as a criterion yielded the most discrepant results for both E7 and E8 and IRT-equating and pre-equating.

Comparing IRT-equating and pre-equating, it can be seen from Table 5 that for E7 pre-equating is actually closer to the criterion equatings but that the composition of the mean squared error is different. For IRT-equating the squared bias is the larger component, whereas for pre-equating the vari-

Table 5
Weighted Mean Squared Difference for Form E7, Using E3, E4, and E5 as the "Old" Form and Three Different Criteria

		IRT-equating Old form			Pre-equating Old form		
		E3	E4	E5	E3	E4	E5
Mean Squared Difference Criterion	C1	2.53	.71	.94	2.04	.79	.70
	C2	2.57	.73	.96	2.11	.84	.74
	C3	3.39	1.33	1.43	2.77	1.19	.88
Variance of Difference Criterion	C1	.53	.20	.89	1.37	.74	.64
	C2	.53	.20	.92	1.42	.79	.68
	C3	.95	.58	1.19	1.83	1.06	.73
Squared Bias Criterion	C1	2.00	.51	.05	.67	.05	.06
	C2	2.04	.53	.04	.69	.05	.06
	C3	2.44	.74	.24	.94	.13	.15

C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as an anchor; C3 is based on equi-percentile equating. The weighting function is the formula score frequency distribution for the first national administration of E7.

Table 6
Weighted Mean Squared Difference for Form E8, Using E3, E4, and E5 as the "Old" Form and Three Different Criteria

		IRT-equating Old form			Pre-equating Old form		
		E3	E4	E5	E3	E4	E5
Mean Squared Difference Criterion	C1	3.99	1.22	1.34	7.67	4.00	3.81
	C2	3.84	1.17	1.29	7.53	3.85	3.76
	C3	4.89	1.70	1.89	8.81	4.89	4.68
Variance of Difference Criterion	C1	1.75	.66	.63	2.83	1.73	1.58
	C2	1.75	.69	.66	2.90	1.73	1.67
	C3	2.49	1.07	1.09	3.74	2.48	2.30
Squared Bias Criterion	C1	2.24	.55	.71	4.84	2.27	2.23
	C2	2.10	.48	.63	4.64	2.12	2.09
	C3	2.39	.63	.80	5.07	2.42	2.38

C1 is based on linear observed score equating using SAT-V and SAT-M as anchors; C2 only uses SAT-V as an anchor; C3 is based on equi-percentile equating. The weighting function is the formula score frequency distribution of the first national administration of E8.

ance of the differences is actually the larger component. For E8, however, the squared bias is the larger component for both IRT-equating and pre-equating.

Summary and Conclusions

This investigation was concerned with how well IRT-equating and pre-equating could reproduce the conversion line for two TSWE forms which had been previously equated by conventional observed score equating methods. The approach was to determine first how well TSWE data fit the three-parameter logistic model and then to compare IRT-equating and pre-equating against three criterion equatings so that discrepancies in equating could be traced to more fundamental questions of fit.

The various procedures for investigating fit suggested several violations of the assumptions of the model. At the item level some of the estimated item-on-ability regressions did not fit the data as well when the *b* parameter had been fixed to its estimated value based on a pretest administration. This is important in pre-equating, since presumably in practical application, parameter estimates would be obtained from pretests. However, the fact that the problem was observed on just a few items suggests that the problem may not be too serious.

At the subscore level it was shown that two factors, corresponding to the two TSWE item types, were required to account for the internal structure of the TSWE, thus suggesting a violation of the undimensionality assumption. This is to be expected, and perhaps so long as the nature of multidi-

mensionality is constant across sample-form combinations, no great harm would occur. It so happened, however, that for Form E8 the two-factor model that fit the other forms did not fit as well. Furthermore, the equating parameters derived under conventional procedures were very different for E8 compared to the other forms. This suggests that the internal structure of E8 was somewhat different.

Departures from the models are to be expected with actual data. The important question, and the focus on this study, is whether such departures seriously affect equating. To answer this question IRT-equating and pre-equating were done for TSWE Forms E7 and E8. The results for E8 were disappointing in that large discrepancies were found between the operational conversion line and the IRT-based equatings. However, since the two-factor model that fit all forms did not fit E8, this appears to be the result of the aberrant internal structure of E8 rather than a failure of IRT-equating. In other words, it is not clear that E8 is properly equated even by conventional methods; hence, for E8 the converted scores may not be a good criterion. Therefore, it is wiser to formulate conclusions based on the results for E7 only.

The equating results for E7 were much more favorable, but some consistent discrepancies were observed, including an overestimation of the mean as well as an underestimation of the standard deviation of the distribution of converted scores. A possible explanation for this discrepancy may be found in the transformation of formula scores to scaled scores.

At least two limitations of the transformation procedure are obvious. One is the assumption of a linear relationship between formula scores and number-correct scores, which is false if students with omitted responses are included. A second limitation of the procedure lies in the fact that to put the new form on scale it is necessary to apply A and B conversion parameters based on observed score equating to true formula scores. Unfortunately, it is not obvious what alternative procedure could be devised to put on scale new forms so long as formula scoring was in use. This suggests that IRT-equating and pre-equating would be more accurate if number-correct scoring was used rather than formula scoring. Number-correct scoring is no panacea, however, so long as tests are speeded (and they always will be for at least some students under the usual administration procedures). The problem is that under number-correct scoring it is advantageous to the student to attempt an item even if he or she has to guess. If they actually do so, they will not be responding as a function of their ability and will thus create a violation of the model (Lord, 1980).

As for pre-equating, based on results for E7 there is reason to be optimistic, since the mean squared differences were not consistently higher for pre-equating across old forms or criteria. However, the criteria for evaluating both IRT-equating and pre-equating are not defensible on other than practical grounds. Thus, unless the comparability of IRT- and pre-equating were to change when evaluated against a more adequate criterion, it can reasonably be expected that, as better procedures for linking parameters are developed (see, e.g., Petersen et al., 1981), pre-equating will prove to be a feasible operational procedure.

References

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

Bejar, I. I., & Wingersky, M. S. *An application of item response theory to equating the Test of Standard Written English* (College Board Report 81-8). New York: College Entrance Examination Board, 1981.

Birnbaum, A. Test scores, sufficient statistics, and the information structure of tests. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley, 1968.

Breland, H. M. *A study of college English placement and the test of standard written English* (Project Report 77-1). Princeton NJ: Educational Testing Service, 1976.

Holland, P. W., & Wightman, L. E. Section pre-equating: A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Test equating.* New York: Academic Press, in press.

Indow, T., & Samejima, F. *On the results obtained by the absolute scaling model and the Lord model of the field of intelligence.* Yokohama: Keio University, Hiyoshi Campus, Psychological Laboratory, 1966.

Jöreskog, K. G. Structural analysis of covariance and correlation matrices. *Psychometrika,* 1978, *43,* 443-477.

Lord, F. M. A study of speed factors in tests and academic grades. *Psychometrika,* 1956, *21,* 31-50.

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

Marco, G. L., Petersen, N. S., & Stewart, E. E. A test of adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), *Proceedings of the 1979 Conference on Computerized Adaptive Testing.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Mosteller, F., & Tukey, J. W. *Data analysis and regression.* Reading MA: Addison-Wesley, 1977.

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.

Samejima, F. A comment on Birnbaum's three-parameter logistic model. *Psychometrika,* 1973, *38,* 221-233.

Sörbom, D., & Jöreskog, K. G. COFAMM: *Confirmatory factor analysis with model modification, a FORTRAN IV program.* Chicago: National Educational Resources, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76-6). Princeton NJ: Educational Testing Service, 1976.

Yen, W. M. The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement,* 1980, *17,* 297-311.

Acknowledgments

The authors express their gratitude to Frederic Lord and Gary Marco, who guided various aspects of the project. Also, Linda Cook, Nancy Petersen, and J. B. Sympson offered many substantive and editorial suggestions, which have improved the report. June Stern was very helpful in the early stages of the project in transmitting some of her knowledge about the most obscure details of the tests. Nancy Wright performed many of the "conventional" analyses and provided additional background information.

Author's Address

Send requests for reprints or further information to Isaac I. Bejar, 20-T, Educational Testing Service, Princeton NJ 08541.

Appendix

Transformation of the *b*'s to Put the LOGIST Output from a New Administration on the Same Scale as the Output from an Old Administration

Robust estimates of scale and location were used to determine the slope and intercept of the line relating the *b*'s estimated from two samples of examinees. The estimate of scale for each of the forms is a biweight estimate (see Mosteller & Tukey, 1977). The formulas are

$$b = \text{median of } b\text{'s}, \qquad [A1]$$

$$u_i = \frac{b_i - \tilde{b}}{9(MAD_b)} \qquad [A2]$$

where MAD_b = Median Absolute Deviation = Median $|b_i - \tilde{b}|$,

$$s_b^2 = \frac{\sum_i' (b_i - \tilde{b})^2 (1 - u_i^2)^4}{[\sum_i' (1 - u_i^2)(1 - 5u_i^2)][\sum_i' (1 - u_i^2)(1 - 5u_i^2) - 1]} \qquad [A3]$$

where $\sum_i'$ indicates summation for $u_i^2 \leq 1$.

Let B_i be the b's on the new sample and b_i be the b's on the old sample. The slope of the line relating the b's is

$$m = \frac{s_B}{s_b} \quad . \qquad [A4]$$

Define an xy coordinate system by

$$x = (b + mB) \quad , \qquad [A5]$$

$$y = (-mb + B) \quad . \qquad [A6]$$

Obtain a robust estimate of location separately for x and for y, using the formulas from Mosteller and Tukey (1977, p. 205). Let y^* = median of the y's,

$$w_i = \begin{cases} (1 - (\frac{y_i - y^*}{c(MAD_y)})^2)^2 & \text{when} \quad (\frac{y_i - y^*}{c(MAD_y)})^2 < 1 \\ 0 & \text{otherwise} \end{cases} \qquad [A7]$$

where MAD_y = Median $|y_i - y^*|$, $c = 6$. Compute a new y^* from the formula

$$y^* = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad . \qquad [A8]$$

Iterate through Equations A7 and A8 until the change between two estimates of y^* is less than .0001. Repeat the process for the x's.

Transform these biweight estimates of location in the xy coordinates back to the bB coordinates and require that the line with slope m pass through this point

$$B^* = \frac{mx^* + y^*}{m^2 + 1} \quad . \qquad [A9]$$

$$b^* = \frac{x^* - my^*}{m^2 + 1} \quad , \qquad [A10]$$

The equation for the line that puts the old parameters on the new parameter scale is

$$b^T = mb + B^* - mb^* \quad [A11]$$

and

$$a^T = (1/m)a \quad . \quad [A12]$$

The equation for the line that transforms the new parameters to the old parameter scale is

$$B^T = \frac{B}{m} - \left[\frac{B^* - mb^*}{m} \right] \quad [A13]$$

and

$$A^T = mA \quad [A14]$$

where a and A are the discrimination estimates based on the old and new samples, respectively.

To put the parameters of a new form onto the same scale as an old form, the new form and the old form must be administered to random samples in a new administration. This was done by spiralling. For the new administration the parameters for the old form are re-estimated and the parameters for the new form are estimated, on random samples of equal size. The new form is put onto the scale of the old form in the new administration by setting the means and standard deviations of the abilities equal. This is done in LOGIST by standardizing the abilities to a mean of 0 and a standard deviation of 1 for both forms. Then, the transformation that puts the old form, new sample (Equations A13 and A14) onto scale is applied to the parameters for the new form to put those parameters onto scale.