Ability Metric Transformations Involved in Vertical Equating Under Item Response Theory

Frank B. Baker
University of Wisconsin—Madison

The metric transformations of the ability scales involved in three equating techniques—external anchor test, internal anchor test, and a pooled groups procedure—were investigated. Simulated item response data for two unique tests and a common test were obtained for two groups that differed with respect to mean ability and variability. The obtained metrics for various combinations of groups and tests were transformed to a common metric and then to the underlying ability metric. The results showed that there was reasonable agreement between the transformed obtained metrics and the underlying ability metric. They also showed that the largest errors in the ability score statistics occurred under the external anchor test procedure and the smallest under the pooled procedures. Although the pooled procedure performed well, it was affected by unequal variances in the two groups of examinees.

One of the practical applications of item response theory (IRT) has been test equating (Lord, 1977, 1980). Such procedures rest on two features of the theory: (1) the existence of a metric for the latent trait (ability) and (2) the invariance principle. Under the theory, test equating reduces to finding a linear transformation for positioning tests and groups of examinees along the ability scale. Although many different equating designs exist (see Angoff, 1971; Lord, 1975), the two classifications—horizontal and vertical equating—reflect the basic issues. Horizontal equating is employed when the groups of examinees are considered to be equivalent and the test forms measure at the same ability levels. The intent of such equating is to remove sampling differences in placing the tests and examinees on a common scale. Vertical equating is employed when the groups of examinees differ in ability level and the tests differ in difficulty level. The goal is to position the tests and groups of examinees along a common scale (Slinde & Linn, 1977).

Three different equating paradigms, based on an anchor test approach, can be identified:

1. The external anchor test procedure, in which two or more groups are administered a separate common test and a test unique to each group.
2. The internal anchor test procedure, in which the items in the anchor test and the items unique to a group are administered as a single test to each group of examinees.
3. The pooled procedure (Lord, 1975), in which all items and groups are pooled into a single data set and analyzed.

The first two procedures are widely used (e.g., see Kolen, 1981; Marco, Petersen, & Stewart, 1980; Vale, Maurelli, Gialluca, Weiss, & Ree, 1981). The third procedure has been used by Cook and Eignor (1981) in a complex vertical equating study. Implementing each of these procedures rests on using a computer program to simultaneously estimate the parameters of the test items and the abil-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 3, Summer 1984, pp. 261–271

0146-6216/84/030261-11$1.80

ities of the examinees from the item response data. However, the maximum likelihood estimation procedures involved yield a metric that is indeterminant with respect to its origin and unit of measurement. To resolve this indeterminancy, the origin and unit of measurement must be set to arbitrary values in each application of these procedures. The LOGIST computer program (Wingersky, Barton, & Lord, 1982; Wood, Wingersky, & Lord, 1976) resolves this issue by standardizing the obtained ability estimates for a given group of examinees to a mean of zero and unit variance (i.e., the scale origin is set to zero and the unit of measurement, $S_{\hat{\theta}}$, to one). In the context of equating, this means that each pairing of a group of examinees and a test yields an ability metric that depends on the distribution of the examinees over the underlying ability metric. Thus, a major task of equating is transforming the several unique group metrics to a common metric.

The present study investigated the transformation of the ability scale metrics involved in vertical equating based on the anchor test approach. Two issues were of interest: (1) The nature of the metric transformations needed under the three equating procedures when the standardization imposed by the LOGIST program is involved, and (2) the impact on the metric transformations of the accuracy with which the parameters of the linking items are estimated under the three equating procedures. A simulation approach was employed in order to use the underlying ability metric as the criterion against which to evaluate the results.

Method and Results

Data Generation Procedures

The underlying ability metric, denoted by θ, was defined as having a midpoint score of zero and a unit of measurement of one. Two normally distributed groups of simulated examinees, each of size 1,000, were created that differed in mean ability level and variability. The parameters of Group 1 (G_1) were $\bar{\theta} = -.253$ and $\sigma_\theta = 1.487$, while Group 2 (G_2) had $\bar{\theta} = .499$ and $\sigma_\theta = .746$, which resulted in groups having different average abilities and variances. Although most of the literature on equating deals with either a one- or three-parameter item characteristic curve model, the two-parameter normal ogive model was employed here. This choice of model was based on two considerations. First, the guessing parameter c of the three-parameter model is not involved in the metric transformation equations (Stocking & Lord, 1983). Second, the parameter c is difficult to estimate, and errors in its estimates are reflected in the estimates of the difficulty parameter (Thissen & Wainer, 1982).

Three simulated tests were established whose item parameter values were expressed in the θ metric. Test U1 consisted of 43 items having $\bar{\alpha} = .3146$ and $\bar{\beta} = -.1835$, where α is item discrimination and β is item difficulty, and was appropriate for Group 1. The items in this test were administered only to Group 1. Test U2 was composed of 43 items with $\bar{\alpha} = .5901$ and $\bar{\beta} = .3378$, and its difficulty level was appropriate to Group 2. In addition Test U2 had higher discriminations than Test U1. The items in this test were administered only to Group 2. The anchor test, A, contained 14 items characterized by $\bar{\alpha} = .4146$ and $\bar{\beta} = -.0129$ and was administered to both groups. The average difficulty of this anchor test was positioned approximately midway between Test U1 and Test U2. The number of items in the anchor test was chosen so that it was 25% of the total number of items when Tests U1 and A or U2 and A were merged. In addition, it is common practice to use a relatively small number of items when linking tests.

Various combinations of the two groups and the three tests defined the data sets used below to examine the three vertical equating paradigms. Table 1 presents the values of the parameters entered into the GENIRV computer program (Baker, 1978) used to generate the four sets of item response data. Each of these data sets was analyzed using the LOGIST program, and the mean values of the obtained item parameter estimates ($\bar{\hat{a}}$, $\bar{\hat{b}}$) are reported in Table 2 by group, test, and equating procedure.

Because of the standardization imposed on the ability estimates by LOGIST, the obtained values of $\hat{a}$, $\hat{b}$ are not in the metric of the item parameters (α, β) used to generate the item response data. To make comparisons, it is useful to express the mean

Table 1
Values of Parameters Used to Generate the Item Response Data

Test	Item Parameters				
	n	$\bar{\alpha}$	σ_α	$\bar{\beta}$	σ_β
U1	43	.3246	.0707	-.1835	.5414
U2	43	.5901	.2045	.3378	.4073
Anchor	14	.4146	.1505	-.0129	.7320
	Ability				
Group	N	$\bar{\theta}$	σ_θ		
G1	1000	-.253	1.487		
G2	1000	.499	.746		
G1 + G2	2000	.123	1.236		

values of the underlying item parameters in the metric resulting from the standardization. Therefore, the mean values of the standardized item parameters ($\bar{a}$, $\bar{b}$) for each test and group combination are reported in Table 2. It should be noted that the numerical values of $\bar{a}$ for Tests U1 and U2 are nearly the same despite the values of $\bar{\alpha}_1$ and $\bar{\alpha}_2$ being quite different in the underlying θ metric.

Table 2
Mean Values of Observed and Standardized Item Parameters by Group and Test

Group	Test	n	Observed $\bar{a}$	Standardized Parameters $\bar{a}$	Observed $\bar{b}$	Standardized Parameters $\bar{b}$
External Anchor Test						
G1	U1	43	.4955	.4678	.0455	.0467
	A	14	.7307	.6165	.1964	.1615
G2	U2	43	.4709	.4402	-.1681	-.2161
	A	14	.3836	.3093	-.7886	-.6862
Internal Anchor Test						
G1	U1A	57	.5249	.5043	.0784	.0749
	Item subset					
	U1	43	.4781	.4678	.0451	.0467
	A	14	.6686	.6165	.1807	.1615
G2	U2A	57	.4182	.4081	-.3249	-.3315
	Item subset					
	U2	43	.4591	.4402	-.1684	-.2161
	A	14	.2929	.3093	-.8057	-.6862
Pooled Tests and Groups						
G1+G2	U1AU2	100	.5277	.5526	-.0703	-.0473
	Item subset					
	U1	43	.3925	.3884	-.2542	-.2480
	U2	43	.6689	.7294	.1348	.1738
	A	14	.5108	.5124	-.1364	-.1100

Vertical Equating

Method. The basic element in vertical equating is the use of a common set of items—an anchor test— that is administered to two or more groups of examinees. When this is done, the differences in ability levels of the groups are reflected in the values of the item difficulty indices, and the differences in group variability are embedded in the values of the item discrimination indices yielded by the several groups for the common items. Thus, the key to transforming the ability scores of one group to the ability metric of another group is the values of the several sets of item parameters. The metric transformation equation given below is based on the mean parameter values; it is due to Haebara (1979) and has been employed by Loyd and Hoover (1980).

$$\bar{\theta}_J = \frac{\bar{\alpha}_K}{\bar{\alpha}_J}(\bar{\theta}_K) + \left(\bar{\beta}_J - \frac{\bar{\alpha}_K}{\bar{\alpha}_J}\bar{\beta}_K\right) \quad , \tag{1}$$

where J and K denote specific group-test combinations.

The parameters in Equation 1 can be replaced by observed or standardized values depending on the context. The relationship between the group variabilities and the mean item discrimination in the tests is given by

$$\bar{\alpha}_J/\sigma_J = \bar{\alpha}_K/\sigma_K \quad . \tag{2}$$

(For a detailed presentation of using Equations 1 and 2 in metric transformations see Baker, 1983.)

To employ Equations 1 and 2 in the present context, it needs to be recognized that the LOGIST program yields a unique metric for each combination of test and group of examinees. Thus, the ability estimates obtained for, say, Group 2 on Test A can be transformed to the metric yielded by the Group 1-Test A combination using the two sets of item parameter estimates for Test A in Equation 1. Due to the use of simulation in the present study, the values of the parameters used to generate the item response data for each group-test combination were known. As a result, it was also possible to transform the theoretical metrics in parallel with the obtained results. Although each examinee's ability score could be transformed, it is more informative to transform the mean and standard deviation of such scores from one metric to another. Due to the parallelism, the obtained ability score statistics can be compared with those based on the parameters and the adequacy of the transformation can be evaluated.

Results. To illustrate the use of Equations 1 and 2, the mean and standard deviation of the ability scores of Group 2, based on the Anchor Test A, were transformed to the metric, denoted by θ_1, based on the Group 1-Test A combination. The transformation was first performed using the standardized parameter values for these combinations from Table 2. Substituting in Equation 1 yielded $\bar{\theta}_1 = .5058$ and substituting in Equation 2 yielded $\sigma_{\theta_1} = .5017$ as the theoretical values of the summary statistics for Group 2-Test A ability scores expressed in the metric that LOGIST would yield for the Group 1-Test A combination. Using the corresponding observed values from Table 2 yielded $\hat{\bar{\theta}}_1 = .5904$ and $S_{\hat{\theta}_1} = .5250$. The error in the Group 2 mean, expressed as a proportion of the theoretical value $[(\hat{\bar{\theta}}_1 - \bar{\theta}_1)/\bar{\theta}_1] \times 100$, was 17%, whereas the standard deviation had an error of 5%.

In the present study, the values of $\bar{\alpha}$, $\bar{\beta}$ for the Anchor Test A were known. Therefore, the Group 2 results could be transformed from the θ_1 metric to the underlying θ metric. If the transformations are proper, the mean and standard deviation of the Group 2 results should be those reported in Table 1. Any discrepancy can be used to evaluate the adequacy of the multiple transformations. Substituting the anchor test parameter values for $\bar{\alpha}$, $\bar{\beta}$ from Table 1 and for $\bar{a}$, $\bar{b}$ from Table 2 in Equations 1 and 2 yielded $\bar{\theta} = .499$ and $\sigma_\theta = .746$, which are exactly the values given for Group 2 in Table 1. Thus, using parameters, the Group 2-Test A results can be transformed correctly to the θ_1 metric and then to the θ metric. Substituting the corresponding observed values $\hat{\bar{a}}$, $\hat{\bar{b}}$ from Table 2 and $\bar{\alpha}$, $\bar{\beta}$ from Table 1 in Equations 1 and 2 yielded $\hat{\bar{\theta}} = .6815$ and $S_{\hat{\theta}} = .925$ for the Group 2-Test A results in the θ metric. The observed values for the Group 2-Test A results expressed in the θ metric differed from the true values, with the mean having an error of $\hat{\bar{\theta}} - \bar{\theta} = .182$ and the standard deviation having an error of $S_{\hat{\theta}} - \sigma_\theta = .179$. A part of

these differences can be attributed to the compounding of estimation errors in the values of $\hat{a}$, $\hat{b}$ yielded for the 14-item common test when the Group 1 values were used a second time in Equation 1.

This example illustrates the dependence of the metric transformation process on the item parameters and their estimates. Because of this, particular attention was paid to the mean values of the item parameters in the several metrics and to the accuracy of the item parameter estimates. An additional reason for attending to the item parameters is that under vertical equating, both the tests and the groups of examinees are to be placed along a common ability scale and the mean item difficulty locates a test on the scale.

External Anchor Test

Method. The logic of the external anchor test procedure for vertical equating can be demonstrated using the groups and tests defined in Table 1. Group 1 takes Tests U1 and A; Group 2 takes Tests U2 and A. The goal is to place the three tests and two groups along a common ability metric. This is accomplished using the item parameter estimates yielded by the common test taken by both groups. To implement this, the four LOGIST analyses (G1U1, G1A, G2U2, and G2A) reported in Table 2 were employed. Four separate metrics could be conceived here, but due to the standardization employed by LOGIST, only two obtained metrics can be distinguished. Both the G1U1 and G1A analyses estimate the ability of Group 1 and yielded a mean of zero and a unit variance for the ability estimates. They differ in the precision with which the abilities are estimated, but from a metric definition point of view they are indistinguishable. This ability metric is called the θ_1 metric.

Similarly, G2U2 and G2A yielded a metric based on Group 2, which is called the θ_2 metric. Because the anchor test is external to Tests U1 and U2, these two tests play no role in defining the metric transformation. As a result, the transformation of ability scores from, say, the θ_2 metric to the θ_1 metric is exactly that given above for the common test situation. Since the item parameters of Test U1 are already in the θ_1 metric, all that remains is to transform the item parameters for Test U2 into the θ_1 metric. This was done for both the standardized parameters and the observed values reported for Test U2 in Table 2, and the results are reported in Table 3.

Results. Using standardized parameters in Equation 2 yielded $\bar{a} = .4402(1)/(.5017) = .8775$, which is the average item discrimination of Test U2 based on Group 2, but expressed in the θ_1 metric yielded by LOGIST for the anchor test taken by Group 1. Note that σ_{θ_j} in Equation 2 was set equal to the value of the standard deviation of ability for Group 2 expressed in the θ_1 metric found in the common test example presented above. The mean item difficulty of Test U2 relative to the anchor test mean difficulty is given by $|\bar{b}_2 - \bar{b}_A| = |-.2161 - (-.6862)| = .4701$ in the θ_2 metric. Then, $|\Delta\bar{b}| = |\bar{b}_2 - \bar{b}_A|\sigma_{\theta_1} = .4701\ (.5017) = .2358$ locates Test U2 relative to the anchor test difficulty in the θ_1 metric. However, the mean ability of Group 1 in the underlying θ metric is well below the mean difficulty of the anchor test. Thus, the mean difficulty of Test U2 relative to the mean ability of Group 1 is $\bar{b} = .2358 + .1615 = .3973$ in the θ_1 metric. Keeping track of the appropriate signs can be aided by making a diagram depicting the relative location of the test characteristic curves and the groups along the θ metric.

Using the obtained results for Test U2 from Table 2 in the same manner yielded $\bar{\hat{a}} = .8969$ and $\bar{\hat{b}} = .5222$ (relative to the mean of Group 1) in the θ_1 metric.

At this point, the item parameters and their estimates for all three Tests U1, U2, and A were in a common θ_1 metric and the item difficulties were relative to the mean of Group 1 in units of σ_θ for this group. The error in the mean discrimination index of Test U2, expressed as a proportion of the theoretical value, was 2%, whereas the item difficulty had an error of 31%.

A check on the adequacy of the equating for Test U2 can be achieved by further transforming these results to the underlying θ metric using the value of $\sigma_\theta = 1.487$ for Group 1 given in Table 1. The theoretical values obtained were $\bar{\alpha} = .5901$ and $\bar{\beta} = .3378$ in the underlying θ metric. Both values

are those given for Test U2 in Table 1. The transformed observed values were $\bar{\hat{\alpha}} = .6032$ and $\bar{\hat{\beta}} = .5235$ in the θ metric. The agreement of the observed and theoretical mean values of discrimination is good ($\bar{\hat{\alpha}} - \bar{\alpha} = .013$), whereas that for difficulty is rather poor ($\bar{\hat{\beta}} - \bar{\beta} = .19$).

Internal Anchor Test

Method. Under this procedure, the test administered to a group of examinees consists of a set of items unique to the group, combined with a set of linking items. If two such tests were administered to two groups, then both groups would have responded to the linking items that serve as an internal anchor test. The mean item parameter estimates yielded by the two groups of examinees for the common items reflect group differences. Consequently, these estimates can be used to transform the metric of one group to that of the other. The nature of the metric transformations involved in this procedure is illustrated using the same items and groups as reported in Table 1.

The 43 items of Test U1 were pooled with the 14 items of the Anchor Test A to form a 57-item test denoted by U1A. The pooled test had $\bar{\alpha} = .3392$ and $\bar{\beta} = -.1416$. The simulated item response data for the administration of this test to Group 1 were formed by pooling the existing results for Tests U1 and A. A second test of 57 items was defined by pooling the items and corresponding item response data for Tests U2 and A, and was denoted by U2A. This test had $\bar{\alpha} = .5470$ and $\bar{\beta} = .2517$. These two data sets were analyzed by LOGIST, and the observed values of $\hat{a}$, $\hat{b}$ are reported in Table 2. It should be noted that the U1A results are in a θ_1 metric and the U2A results are in a θ_2 metric. In order to transform the metrics, the mean values of item difficulty and discrimination for the linking items are needed. The underlying values of $\bar{\alpha} = .4146$ and $\bar{\beta} = -.0129$ are those given for Test A in Table 1. The standardized values resulting from the use of LOGIST and the observed values are reported as Item Subset A under Tests U1A and U2A in Table 2.

Results. To illustrate the metric transformations for the internal anchor test procedure, the results for Group 2 again were transformed to a θ_1 metric. The transformation of the summary statistics for ability are given first, followed by those for the item parameter estimates.

The underlying values, the values resulting from the LOGIST standardization for the mean and standard deviation of the ability scores, and the item parameters of the anchor test for Group 2, are the same as those reported for the previous two examples. The theoretical values yielded by Equations 1 and 2 for Group 2 in the θ_1 metric were again $\bar{\theta}_1 = .5058$ and $\sigma_{\theta_1} = .5017$. The observed summary statistics for the Group 2 ability scores can be obtained by substituting $\bar{\hat{a}}_K = .2929$, $\bar{\hat{b}}_K = -.8057$ and $\bar{\hat{a}}_J = .6686$, $\bar{\hat{b}}_J = .1807$ from Table 2 into Equation 1, which yielded $\hat{\bar{\theta}}_1 = .5337$. From Equation 2, $S_{\hat{\theta}_1} = .4381$ is obtained.

At this point, both the theoretical and observed summary statistics for the Group 2 ability scores based on the anchor test were expressed in the metric yielded by LOGIST for the Group 1-Test A results. The mean ability of Group 2 differed from the theoretical value by 6% and the error in the standard deviation was 13%. Compared to the external anchor test results, the mean ability had a smaller error but the standard deviation had a larger error. Since the values of $\bar{\alpha}$ and $\bar{\beta}$ were known, the summary statistics for the Group 2 ability scores could be transformed from the θ_1 metric to the underlying θ metric using Equations 1 and 2. As anticipated, the theoretical values obtained are those reported in Table 1 for Group 2. The observed values of the transformed mean ($\hat{\bar{\theta}} = .5566$) and standard deviation ($S_{\hat{\theta}} = .7065$) in the θ metric differed only slightly ($|\hat{\bar{\theta}} - \bar{\theta}| = .055$, $|S_{\hat{\theta}} - \sigma_\theta| = .040$) from the true values.

The item parameter estimates for the U1 subset of items were already in a θ_1 metric; hence, only the U2 estimates remained to be transformed. This is done below for both the standardized parameters and the observed values of Table 2 and the results are reported in Table 3.

The theoretical values for Test U2 are the same as for the external anchor test case given above. The transformed observed results for Test U2 were $\bar{\hat{a}} = 1.0479$ and $\Delta\bar{\hat{b}} = .2792$ in a θ_1 metric, but $\Delta\bar{\hat{b}}$

Table 3
Summary Statistics for Test U2 and Group 2 by Metric

External Anchor	θ_2 metric		Transformed θ_1 metric		Transformed θ metric	
	Observed	Theoretical	Observed	Theoretical	Observed	Theoretical
$\bar{\alpha}$	.4709	.4402	.8969	.8775	.6032	.5901
$\bar{\beta}$	-.1681	-.2161	.5222	.3973	.5235	.3378
$\bar{\theta}$	0	0	.5904	.5058	.6815	.4990
σ_θ	1	1	.5250	.5017	.9250	.7460
Internal Anchor						
$\bar{\alpha}$	.4591	.4402	1.0479	.8775	.7047	.5901
$\bar{\beta}$	-.1684	-.2161	.4599	.3973	.4309	.3378
$\bar{\theta}$	0	0	.5337	.5058	.5566	.4990
σ_θ	1	1	.4381	.5017	.7065	.7460

is relative to the anchor test mean difficulty rather than Group 1 mean ability. Adjusting for this difference yielded $\bar{b} = .4599$. The error in the mean discrimination index was 19% and the error in the mean item difficulty was 16%.

A check on the adequacy of the equating for Test U2 can be achieved by transforming the above results to the θ metric via $\sigma_\theta = 1.487$ for Group 1. The theoretical values are the same as those reported for the external anchor test case and are the values given in Table 3. The observed values are $\bar{\hat{\alpha}} = .7047$ and $\bar{\hat{\beta}} = .4309$. The agreement of the observed and theoretical mean values is not particularly good ($\Delta = .10$) for either index.

Pooled Procedure

Method. A rather ingenious procedure for placing several tests and groups on a common metric using a set of linking items is due to Lord (1975). The test administered to a given group of examinees consists of the items unique to the group plus the common set of linking items. Thus, the basic test definitions and administration to groups are the same as for the preceding internal anchor test example. For parameter estimation purposes, the item response data for Group 1 on Test U1A were pooled with that for Group 2 on Test U2A, resulting in a 100-item test and a group of 2,000 examinees. From a procedural point of view for the LOGIST program, the U2 items are considered as not reached by the Group 1 examinees. The U1 items are treated as not reached by Group 2. The items in Test A are responded to by both groups. Part of the rationale is that the common items are responded to by twice as many examinees and will play a larger role in determining the obtained ability metric. Since all item and ability parameters are estimated in a single LOGIST analysis, the three tests and two groups will share a common ability metric.

In terms of the underlying parameters used to generate the item response data, the resultant 100-item test had $\bar{\alpha} = .4471$ and $\bar{\beta} = .0645$. The pooled group of examinees had $\bar{\theta} = .123$ and $\sigma_\theta = 1.236$. The composite item response data were then analyzed using LOGIST. Table 2 reports the summary statistics of the item parameter estimates for the pooled data as well as separately for each subset of items.

Results. The obtained, underlying, and standardized values for the ability summary statistics are reported in Table 4. In general the agreement between the observed and standardized values was quite good.

Because the primary interest was in placing the three tests and the two groups of examinees on a common scale, the obtained results were transformed to the θ scale and are reported in Tables 4 and 5 so that comparisons with the underlying parameter values can be made.

The transformation of the obtained ability scores (Table 4) to the θ metric uses the item parameter estimates for the linking items from Table 2 and the corresponding parameter values from Table 1. The transformed summary statistics for ability obtained from Equations 1 and 2 were as follows:

pooled groups, $\bar{\hat{\theta}} = .1551$ and $S_{\hat{\theta}} = 1.232$;
Group 1, $\bar{\hat{\theta}} = -.1657$ and $S_{\hat{\theta}} = 1.4461$;
Group 2, $\bar{\hat{\theta}} = .4694$ and $S_{\hat{\theta}} = .8814$.

The differences between the transformed estimated mean ability and the underlying values were .03, .09, and .03 for the pooled group, Group 1, and Group 2, respectively. The differences between the transformed obtained standard deviation of the ability scores and underlying values in the θ metric were 0.0, −.04, and .14 for the pooled group, Group 1, and Group 2, respectively. Only the difference for Group 2 appears to be discrepant.

The mean item parameter estimates for the several tests were transformed to the θ metric and are reported in Table 5. The values of $\bar{\hat{\alpha}}$ were within .019 of $\bar{\alpha}$, except for Test U2, which yielded $\bar{\hat{\alpha}} - \bar{\alpha} = .049$. A similar pattern held for the mean item difficulties with the exception of Test U2, which yielded $\bar{\hat{\beta}} - \bar{\beta} = .048$. The anchor test results were very close to the underlying values ($\bar{\hat{\alpha}} - \bar{\alpha} = .001$, $\bar{\hat{\beta}} - \bar{\beta} = .032$). It is of interest that such accurate values were obtained with only 14 items; however, these item parameter estimates were based on 2,000 examinees, which should enhance their accuracy.

Discussion

Under the external anchor test procedure, the 14 linking items constituted a separate test administered to each of the groups. Merging the linking items with each of the groups' unique items resulted in the instruments administered to each group in the internal anchor test procedure. Merging the item response data for these two instruments was the basis of the pooled procedure. Because of this sequence, the metric transformations would be ex-

Table 4
Summary Statistics for Ability Based Upon Pooled Tests and Pooled Groups

		Group		
		G1, G2 Pooled	G1	G2
Mean				
	Underlying	.1230	-.2530	.4990
	Standardized	0.0	-.3043	.3042
	Observed	.0001	-.2604	.2551
	Transformed	.1551	-.1657	.4694
Standard Deviation				
	Underlying	1.2360	1.4870	.7460
	Standardized	1.000	1.2030	.6036
	Observed	1.0001	1.1701	.7131
	Transformed	1.232	1.4461	.8814
N		1979	979*	1000

* 21 cases lost due to null or perfect scores.

Table 5
Transformed Summary Statistics for Item Parameters and Estimates in the θ Metric Under the Pooled Procedure

		Test			
		Pooled	Item Subsets		
			U1	U2	A
Theoretical	$\bar{\alpha}$	.4471	.3246	.5901	.4146
	$\bar{\beta}$	.0645	-.1835	.3378	-.0129
Observed	$\bar{\alpha}^*$	.4269	.3176	.5412	.4132
	$\bar{\beta}^*$	-.036	-.1887	.2896	-.0456

pected to yield the largest errors in the ability estimate statistics under the external anchor test procedure and the smallest under the pooled procedure. Verifying this pattern of results was obscured by the standardization used by the LOGIST program to remove the metric indeterminancy inherent in the maximum likelihood estimation procedures.

To work around this, the mean and standard deviation of the ability estimates for Group 2 under the first two procedures were first transformed to the θ_1 metric, based on the Group 1 results, and then transformed to the underlying θ metric. The results for the pooled group were transformed directly to the θ metric. The transformed results were then compared with the parameter values used to generate the item response data. The values of $|\hat{\theta} - \bar{\theta}|$ showed the anticipated progression, yielding values of .18, .06, and .03 for the three procedures, respectively. The value of $|S_{\hat{\theta}} - \sigma_\theta|$ was large, .18 and .14 for the anchor test and pooled procedures, respectively, but small, .04, for the internal anchor test procedure.

Since in practice an obtained metric can only be transformed into another obtained metric, comparison of the results in the θ_1 metric also was of interest. The errors in the mean ability estimates for Group 2 in the θ_1 metric were 17% and 6% of their theoretical values for the external and internal anchor test procedures, respectively. The corresponding errors in the standard deviations were 5% and 13%. The Group 2 results under the pooled procedure can be compared with those expected under the LOGIST standardization based on the pooled group metric rather than a θ_1 metric. The error in the mean ability estimate was .05 and the error in the standard deviation was .11 or 10% and 15%, respectively.

In the case of the item parameter estimates, the same pattern of error would be expected as for the ability estimates, across the sequence of equating procedures. Although the number of linking items is the same in all cases, better ability estimates, due to longer tests, should improve the item parameter estimates. When compared to the theoretical values expected under the LOGIST standardizations involved, the errors in $\bar{\hat{a}}$ and $\bar{\hat{b}}$ for the anchor test results reported in Table 2 behaved differently in the two groups. In the case of $\bar{\hat{a}}$ based on Group 1, the error decreased from 19% to 8%, and $\bar{\hat{a}}$ based on Group 2 decreased from 24% to 5% for the external and internal anchor test procedures, respectively. The errors in $\bar{\hat{b}}$ were 21% and 12% for Group 1 and 10% and 17% for Group 2.

When both groups were merged under the pooled procedure, the item parameter estimates of the common items were based on 2,000 examinees, and the errors in $\bar{\hat{\alpha}}$ and $\bar{\hat{b}}$ were .3% and 24% of their theoretical values. The relative error in $\bar{\hat{b}}$ was large, but the base was near zero and the absolute difference was only .03. When the Test U2 results were transformed from the obtained metric to the θ metric, both $\bar{\hat{\alpha}}$ and $\bar{\hat{\beta}}$ follow the anticipated pattern. The absolute errors in $\bar{\hat{\alpha}}$ were .013, .10, and .05, whereas those for $\bar{\hat{\beta}}$ were .19, .10, and .05 for the external, internal anchor test, and pooled procedures, respectively. The anomalous result is

the very small error (.013) for $\bar{\hat{\alpha}}$ yielded by the external anchor test procedure. In the case of the external and internal anchor test procedures, the U2 results in the θ_1 metric can be compared with the theoretical results under the LOGIST standardization. The proportional errors for $\bar{\hat{a}}$ and $\bar{\hat{b}}$ were 2% and 29%, and 31% and 16% for the external and internal anchor test procedures, respectively. When the Test U2 results in the pooled group metric were transformed to the θ metric, the pooled procedure yielded errors of 8% and 22% for $\bar{\hat{a}}$ and $\bar{\hat{b}}$, respectively. Since the metric transformation equation employed was based on the mean values of the item parameter estimates of the linking items, accuracy is important.

The metric transformations involved in the external and internal anchor test procedures were the same at the theoretical level, and differences between the observed and theoretical results were associated with parameter estimation errors. When the Group 2 results were compared at the θ metric level, the results yielded by the internal anchor test procedure were better than those due to the external anchor test procedure. Against this same criterion, the pooled procedure was the best with respect to item discrimination and mean ability. However, it yielded somewhat poorer results for item difficulty and the variability of ability. Based on the present results, the pooled procedure appears to have problems coping with merged groups that have different underlying variabilities.

A feature of the vertical equating results presented above was that all of the analyses were performed using the same sets of item response data. Under the invariance principle, the item parameters should be group independent, the ability parameters item independent, and both should be invariant with respect to the equating design. However, this was not the case since the obtained numerical values varied widely and the estimation errors were somewhat inconsistent across techniques. Part of this problem can be attributed to the maximum likelihood estimation technique and part to the standardization imposed by the LOGIST computer program. For example, under the internal anchor test procedure, Group 1 yielded $\bar{\hat{a}} = .6686$, $\bar{\hat{b}} = .1807$ and Group 2 yielded $\bar{\hat{a}} = .2929$, $\bar{\hat{b}} = -.8057$ for the 14-item anchor test.

Despite the items being the same, these results were quite discrepant in terms of their numerical values. In the case of Group 2, many values of $\hat{a}$ were quite small, and when this holds, the maximum likelihood estimation methods can have problems in obtaining precise estimates of the item difficulty parameter. Since the transformation employed here (Equation 1) depended on $\bar{\hat{a}}$ and $\bar{\hat{b}}$, problems in their estimation get carried into the resultant ability scale metric. Although the errors in $\bar{\hat{a}}$ were generally small, the errors in $\bar{\hat{b}}$ were generally larger, and the latter probably can be attributed to the artificially low numerical values of $\bar{\hat{a}}$ for Group 2 resulting from the LOGIST standardization.

The influence of the level of precision in the item parameter estimates of the linking items on equating procedures has been recognized, and techniques are appearing to take this factor into account (see Bejar & Wingersky, 1981; Haebara, 1981; Stocking & Lord, 1983). The problems arising from the standardization imposed by LOGIST on the pooled group results can best be handled by using groups of equal underlying variability.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington DC: American Council on Education, 508–600.

Baker, F. B. (1978). *GENIRV: A program to generate item response vectors* [Computer Program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Baker, F. B. (1983). Comparison of ability metrics obtained under two latent trait procedures. *Applied Psychological Measurement, 7*, 97–110.

Bejar, I., & Wingersky, M. S. (1981). *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.

Cook, L. L., & Eignor, D. R. (1981). *Score equating and item response theory: Some practical considerations*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles CA.

Haebara, T. A. (1979). *Method for investigating item bias using Birnbaum's three-parameter logistic model*

(Iowa Testing Programs Occasional Paper Number 25). Iowa City IA: University of Iowa, Iowa Testing Programs.

Haebara, T. A. (1981). *Least squares method for equating logistic ability scales: A general approach and evaluation* (Iowa Testing Programs Occasional Paper Number 30). Iowa City IA: University of Iowa, Iowa Testing Programs.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18,* 1–11.

Lord, F. M. (1975). *A survey of equating methods based upon item characteristic curve theory* (RB-75-13). Princeton NJ: Educational Testing Service.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117–138.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 169–194.

Marco, G. L., Peterson, N. S., & Stewart, E. E. (1980). A test of the adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computer Adaptive Testing Conference* (pp. 167–196). Minneapolis MN: University of Minnesota, Department of Psychology.

Slinde, J. A., & Linn, R. L. (1977). Vertical equated tests: Fact or phantom? *Journal of Educational Measurement, 14,* 23–32.

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

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, J. M. (1981). *Methods for linking item parameters* (Report APHRL-TR-81-10). Brooks Air Force Base TX: Air Force Human Resources Laboratory (AFSC).

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST User's Guide*. Princeton NJ: Educational Testing Service.

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

Author's Address

Send requests for reprints or further information to Frank B. Baker, Department of Educational Psychology, University of Wisconsin, Madison WI 53706, U.S.A.