Comparison of Ability Metrics Obtained under Two Latent Trait Theory Procedures

Frank B. Baker
University of Wisconsin

Both the BICAL and LOGIST computer programs implement a maximum likelihood procedure for jointly estimating the item and ability parameters. The two programs differ, however, with respect to (1) the anchoring procedures used to overcome the metric indeterminancy of the paradigm; (2) the item characteristic curve models employed; and (3) how the examinees are grouped within the estimation process. Three simulated sets of item response data based upon a known underlying ability metric were used to investigate the metric recovery capabilities of the two computer programs. The results showed that both programs recovered a transformation of the underlying metric via a common equation, but the elements used in this equation were program specific. The transformation of the metric yielded by BICAL to the underlying metric depended only upon the item characteristic curve parameters, whereas the LOGIST transformation also depended upon the frequency distribution of the estimated ability scores over the underlying ability metric. The empirical results indicate that both transformations are quite sensitive to errors in the average value of the obtained item discrimination indices. Because LOGIST groups examinees by ability levels and BICAL does so by raw score levels, the variability of the transformed ability estimates yielded by BICAL were smaller than those from LOGIST. The results suggest that when comparing results yielded by the two computer programs, particular attention should be paid to the characteristics of the obtained metrics.

A number of years ago Uhr (1963) predicted that future psychological theories will be defined in terms of computer programs. In the case of latent trait/item response theory, this is already the case to some degree. The computer program implementations of the Birnbaum (1968) paradigm for jointly estimating item and examinee parameters via maximum likelihood procedures are the operational definitions of this facet of the theory. This estimation process attempts to recover the metric of the ability scale underlying a particular measuring instrument from the information in the examinees' item responses. However, the ability metric yielded by the Birnbaum paradigm is indeterminant with respect to its location and unit of measurement. As a result, computer programs must employ "anchoring" procedures to remove this indeterminancy.

Two- and three-parameter item characteristic curve models (normal and logistic) lead to an obtained ability metric that must have both its location and unit of measurement set to some value via an anchoring procedure. The one-parameter logistic (Rasch) model yields an ability metric that must be anchored only with respect to its location. The choice of how this anchoring is accomplished becomes a decision made by the computer program developer. One consequence of the choice of specific anchoring procedures is

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 1, Winter 1983, pp. 97–110

0146-6216/83/010097-14$1.70

that the numerical values of the parameter estimates produced by one program are not directly comparable with those yielded by another program. For example, in a set of recent papers (Forsyth, Saisangjan, & Gilmer, 1981; Guskey, 1981; Yen, 1981) the LOGIST program (Wood, Wingersky, & Lord, 1976) was used to estimate the item difficulty and ability parameters under a one-parameter logistic (Rasch) model; however, this approach will not yield the same numerical values of the difficulty and ability estimates as would BICAL (Wright & Mead, 1977) for the same data set.

An important feature of the Rasch model is that there is a mathematical requirement for the item discrimination parameter to be unity for all items in the test. This is the basis for setting the unit of the ability metric equal to one. Conventional wisdom is that this assumption is met when all items share a common value of the item discrimination parameter (see Hambleton & Cook, 1977; Yen, 1981), but among measurement specialists it is also understood that any discrepancy between this common value and one is compensated for by adjusting the unit of measurement of the ability metric (see Wood, 1978). Yet, this compensation is rarely mentioned in the literature dealing with the Rasch model; for example, it is not mentioned in Wright (1968, 1977) or Wright and Stone (1979). However, numerous studies have dealt with the effect of item discrimination parameter values upon the goodness of fit of the Rasch model to item response data (see, for example, Dinero & Hartel, 1977; Forsyth et al., 1981); but none have dealt with the manner in which the violation of the mathematical requirement of a unit value of the item discrimination parameter affects the obtained metric.

The two most widely used computer programs based upon latent trait/item response theory are LOGIST and BICAL, which take quite different approaches to anchoring the ability scale metric. Because of this, the present study investigated the degree to which each program recovers a known underlying metric. In addition, particular attention was paid to the role of the item discrimination parameters in determining the metric yielded under the Rasch model. Procedures were employed that enabled the metric yielded by each program to be expressed in terms of the underlying metric. Thus, the metric recovery capabilities of each program could be evaluated and appropriate comparisons made between the two sets of results. Basically, the problem of interest was one of equating among (1) the true underlying metric, (2) the metric yielded by LOGIST, and (3) the metric yielded by BICAL for a common data set.

Method

LOGIST and BICAL Metric Procedures

The item characteristic curve (ICC) model underlying the LOGIST computer program is the normal ogive; but to reduce the computational load, the logistic model is used to approximate the normal ogive. Although the three-parameter model is commonly used when LOGIST is employed, the two-parameter logistic model was employed here because the usual Rasch model does not encompass guessing. Under the two-parameter logistic model, the probability of a correct response is given by

$$P(u_{ij} = 1 \mid \theta_j, \alpha_i, \beta_i) = \frac{1}{1 + e^{-1.702\,\alpha_i(\theta_j - \beta_i)}} \qquad [1]$$

where

u_{ij} = the response of examinee j ($j = 1, 2, \ldots, M$) to item i ($i = 1, 2, \ldots, n$), and $u_{ij} = 1$ for a correct response and zero otherwise;
θ_j = the ability level of examinee j;
α_i = the item discrimination parameter $-\infty \leq \alpha_i \leq \infty$;
β_i = the item difficulty parameter $-\infty \leq \beta_i \leq \infty$; and
1.702 = a factor used to obtain agreement between the normal and logistic ogive.

The ICC underlying the BICAL program is the one-parameter logistic model given by

$$P(u_{ij}|\theta_j,\beta_i) = \frac{1}{1 + e^{-1(\theta_j - \beta_i)}} \quad [2]$$

which is the two-parameter logistic with α_i set equal to unity.

It should be noted that the LOGIST program reports the values of α_i as if they were normal ogive parameters, i.e., they do not incorporate the 1.702 factor, while BICAL employs $\alpha_i = 1$ in a logistic frame of reference.

Both the LOGIST and BICAL computer programs use the Birnbaum (1968) maximum likelihood estimation paradigm to jointly estimate the vectors of item and ability parameters from the observed $n \times M$ matrix of dichotomously scored item responses. The LOGIST program anchors the obtained metric by standardizing the ability scores of the M examinees, thus fixing the mean at zero and the unit of measurement to $\sigma_{\hat{\theta}}$. This metric will be called the normal metric and be denoted by θ_N. It should be noted that this anchoring procedure makes the obtained metric a function of the frequency distribution of the examinees' ability scores rather than of the underlying metric alone. The BICAL program anchors the location of the metric by setting the mean value of the estimated item difficulty parameters to zero (Wright & Douglas, 1977). This is accomplished by standardizing the estimated item difficulties using

$$\hat{\beta}_i = \frac{\hat{\beta}'_i - \overline{\hat{\beta}'_i}}{\sigma_{\beta'_i}} \quad [3]$$

but $\sigma_{\beta'_i}$ is set to unity and then

$$\hat{\beta}_i = \hat{\beta}'_i - \frac{\sum_{i=1}^{n} \hat{\beta}'_i}{n} \quad [4]$$

where $\hat{\beta}'_i$ is the value of the item difficulty yielded by the joint estimation procedure. The estimated ability scores are in the same metric as the $\hat{\beta}_i$ and will be denoted by θ_R. The rescaled item difficulty estimates are also multiplied by the factor $[(n - 1)/n]$ within BICAL before they are reported. This compensates for the inclusion of a given item difficulty within the sum of the item difficulties.

There are three caveats that must be recognized in regard to these metric anchoring procedures. First, both BICAL and LOGIST do not employ any external frames of reference when anchoring the underlying metric, and both assume the n items and M examinees constitute the only populations of interest. Second, a crucial difference between BICAL and LOGIST is that the BICAL anchoring process depends on the item discrimination parameters of all items in the test having a value of unity. Because of this restriction, the Rasch model will yield the same ability estimate of all examinees obtaining the same raw score, whereas the LOGIST program will yield a separate, not necessarily unique, ability estimate for each of the $\binom{n}{r}$ item response patterns yielding a test score of r. Third, LOGIST produces a metric corresponding to a normal ogive ICC model, while the BICAL program produces a metric corresponding to a logistic model for the ICC.

Data

In order to compare the metrics yielded by the LOGIST and BICAL programs, three sets of simulated item response data were used in the present study. All of the data sets had a common ability scale with

a midpoint of zero and a unit of measurement of one. This metric will be called the underlying metric and will be denoted by θ_T. A computer program (GENIRV; Baker, 1978) was used to generate item response data, given an examinee's ability level in the θ_T metric and the parameters β_i, α_i for each of the *n* items in the test based upon a normal ogive model for the ICC. This data generation procedure meets the theoretical requirement that the metric be such that the ICC model holds simultaneously for all items in the test (Lord & Novick, 1968). The simulated item response data for each group were then analyzed via BICAL and LOGIST. The basic plan was to transform the obtained metrics to the underlying metric and then to evaluate the metric recovery of the two computer programs. The evaluation vehicle chosen was to calculate the mean and standard deviation of the ability estimates and to express them in the underlying metric. If a proper transformation was achieved, the summary statistics should agree with those based upon the underlying ability scores.

The fundamental equation for performing the metric transformation has been given by Loyd and Hoover (1980) as

$$\theta_1 = \frac{\bar{\alpha}_2}{\bar{\alpha}_1}\theta_2 + \left[\bar{\beta}_1 - \frac{\bar{\alpha}_2}{\bar{\alpha}_1}\bar{\beta}_2\right] \qquad [5]$$

and it holds in the parameters. When estimates yielded by different computer programs are involved, Equation 5 does not take the anchoring procedures into account. Since both BICAL and LOGIST use different anchoring procedures, it is useful to express the ability and difficulty parameters in terms of the resultant units of measurement. Thus, the numerical values of θ_1, β_1 are in units of D_1 and the values of θ_2, β_2 are in units of D_2. Introducing these units of measurement in Equation 5 yields a general metric transformation formula of

$$\theta_1 = \frac{\bar{\alpha}_2}{\bar{\alpha}_1}(\theta_2)\frac{D_2}{D_1} + \left[\bar{\beta}_1 - \frac{\bar{\alpha}_2}{\bar{\alpha}_1}(\bar{\beta}_2)\frac{D_2}{D_1}\right] \qquad [6]$$

which is a linear function. Using Equation 6, θ_R can be transformed to θ_T, and θ_N to θ_T; then the results yielded by BICAL and LOGIST can be compared easily.

Three sets of simulated item response data were used to illustrate the facets of the metric recovery process that were of interest. The first data set demonstrated the basic issues involved in the metric transformation process and, in particular, the impact of the average value of α_i upon the obtained Rasch metric. The second set of data illustrated the relative nature of the location of the mean estimated ability scores as compared to the average estimated item difficulty under the two approaches. The third data set involved extreme groups and illustrated the impact of the variability of the underlying ability score distribution and the anchoring procedures upon the obtained ability metric. The last set also involved many of the same issues as are encountered in test equating.

Results

Data Set Based Upon Equivalent Items

The first data set was based upon a population of 1,100 simulated examinees whose ability scores were normally distributed (0, 1) over the θ_T scale. The test was defined as consisting of 20 equivalent items ($\alpha_T = .5$, $\beta_T = 0$) under a normal ogive model for the ICC. The corresponding two-parameter logistic model would have $\alpha = 1.702$, $\alpha_T = .851$, and $\beta = 0$.

The summary statistics of the parameter estimates yielded by BICAL and LOGIST for this data set are reported in Table 1. Both programs eliminated four examinees having raw scores of zero, and a group size of 1,096 resulted. The parameter estimates yielded by LOGIST ($\bar{\hat{\theta}}_N = 0$, $S_{\hat{\theta}N} = 1$; $\bar{\hat{\alpha}}_N = .535$ and $\bar{\hat{\beta}}_N = -.029$) were very close to the underlying specifications. This should be the case since $\sigma_{\theta T} = 1$ optimizes the data with respect to the anchoring procedure used by LOGIST. BICAL yielded parameter estimates ($\bar{\hat{\theta}}_R = .04$, $S_{\hat{\theta}R} = .79$ and $\bar{\hat{\beta}}_R = 0$) that appear similar to the LOGIST results. However, the two sets of parameter estimates are each in their own metric. The difference in metrics is apparent in the range of estimated ability scores (-3.53 to 3.47) and (-2.79 to 2.79) for LOGIST and BICAL, respectively, as well as in the standard deviations (1 and .79) of the ability estimates.

In order to determine if BICAL recovered a function of the underlying metric, the obtained Rasch ability metric $\hat{\theta}_R$ was transformed to the θ_T metric using the known item parameter values. The terms in Equation 6 were defined as follows:

$$\theta_1 = \hat{\theta}_T,\ D_1 = 1,\ \bar{\alpha}_1 = 1.702\bar{\alpha}_T = .851,\ \bar{\beta}_1 = \bar{\beta}_T = 0 \qquad [7]$$

$$\theta_2 = \hat{\theta}_R,\ D_2 = 1,\ \bar{\alpha}_2 = \alpha_R = 1,\ \bar{\beta}_2 = \bar{\hat{\beta}}_R = 0 \qquad [8]$$

and substituting in Equation 6 yields

$$\hat{\theta}_T = \frac{1.0}{.851}(\hat{\theta}_R)\frac{1}{I} + \left[0 - \frac{1.0}{.851}(0)\frac{1}{I}\right] = 1.175\,\hat{\theta}_R = .047 \qquad [9]$$

If in Equation 6, $\theta_1 = \theta_2$, $D_1 = \sigma_1$, $D_2 = \sigma_2$, and $\bar{\beta}_1 = \bar{\beta}_2 = 0$, the following relation results:

$$\bar{\alpha}_1\sigma_1 = \bar{\alpha}_2\sigma_2 \qquad [10]$$

which can be solved for σ_1 yielding

Table 1
Summary Statistics for 20 Equivalent Items ($\alpha_T = .5$, $\beta_T = 0$) and 1096 Examinees N(0,1)

Estimate		LOGIST	BICAL
$\hat{\theta}$	max	3.47	2.79
	mean	0.0	.04
	min	-3.53	-2.79
	$S_{\hat{\theta}}$	1.00	.79
$\hat{\beta}$	max	.100	.096
	mean	-.029	.000
	min	-.135	-.074
	$S_{\hat{\beta}}$	.065	.054
$\hat{\alpha}$	max	.739	
	mean	.525	
	min	.437	
	$S_{\hat{\alpha}}$	.081	

$$\sigma_1 = \frac{\bar{\alpha}_2}{\bar{\alpha}_1}\sigma_2. \qquad [11]$$

An interesting aspect of the BICAL results was that the standard deviation of the ability scores was not used in the anchoring procedure; but Equation 11 can be used to transform $S_{\hat{\theta}R}$ to $S_{\hat{\theta}T}$ and yields

$$S_{\hat{\theta}_T} = 1.175(S_{\hat{\theta}_R}) = 1.175(.79) = .92 \qquad [12]$$

which, though an underestimate, is quite close to the underlying value of 1.0. Similar results can be obtained for $S_{\hat{\theta}N}$ yielded by LOGIST and

$$S_{\hat{\theta}_T} = \frac{.535}{.500}(1.0) = 1.07 \qquad [13]$$

which is also close to the underlying value.

In this data set, the true value of the item discrimination parameter (.851) under a logistic model was less than the a priori value of 1.0 used in the Rasch model. As a result, the BICAL program had to compress the ability metric until the obtained value of $\bar{\hat{\alpha}}_R$ was equal to unity. The scale conversion factor $1/1.702\bar{\alpha}_T$ was effective in compensating for the a priori use of $\alpha_R = 1.0$. The importance of this result is that the BICAL program automatically adjusted the ability scale to compensate for the value of $\bar{\hat{\alpha}}_i$ but gives no indication of having done so. When analyzing a single data set, this is of little importance, but when making comparisons across separate analyses, even with the same test, different metric adjustments can be involved but not detected unless the item discrimination parameters or their estimates are known.

Data Set Based Upon Nonequivalent Items

In the previous example, equivalent items and a unit normal distribution of true ability scores were used to insure that the anchoring procedures used by the two computer programs would not interfere with illustrating the impact of the item discrimination parameters upon the metric recovery. The present example employed a data set that illustrates the relative nature of the item and ability parameter estimates as well as the effect of a low average value of the item discrimination parameters. The true ability scores of the 1,100 simulated examinees were normally distributed over the ability scale with a mean of $\bar{\theta}_T = -.5$ and unit variance. The simulated test consisted of 40 items based upon a normal ogive model. The item discrimination parameters α_{Ti} were selected at random from a uniform distribution having the range .19 to .39 and yielded $\bar{\alpha}_T = .27$, which is a rather low average value. The item difficulty parameters β_{Ti} were selected at random from a normal distribution and yielded $\bar{\beta}_{Ti} = .464$ and a standard deviation of 1.086. The data generation specifications yielded an absolute difference of .964 units between $\bar{\theta}_T$ and $\bar{\beta}_T$.

The GENIRV program was used to generate the binary item responses of the 1,100 examinees based upon a test of 40 items defined by random pairing of the α_{Ti} and β_{Ti}. The generated data were analyzed by both computer programs, and a summary of the parameter estimates is given in Table 2. No simulated examinees were removed for having scores of 0 or 40 by either computer program. BICAL yielded $\bar{\hat{\theta}}_R = -.43$, $S_{\hat{\theta}R} = .44$ and $\bar{\hat{\beta}}_R = 0$. The LOGIST results were $\bar{\hat{\theta}}_N = -.016$, $S_{\hat{\theta}N} = 1.052$, $\bar{\hat{\alpha}}_N = .321$, and $\bar{\hat{\beta}}_N = .830$.

Again, it is useful to transform the BICAL results to the θ_T metric via Equation 6 using the true values of $\bar{\alpha}_T$ and $\bar{\beta}_T$; then

$$\bar{\hat{\theta}}_T = \frac{1.0}{.4595}\bar{\hat{\theta}}_R + \bar{\beta}_T = 2.176(-.43) + .464 = -.472 \qquad [14]$$

Table 2
Summary Statistics for 40 Non-Equivalent Items, 1100 Examinees

Estimate		LOGIST	BICAL
$\hat{\theta}$	max	2.73	1.78
	mean	-.016	-.43
	min	-4.00	-2.55
	$S_{\hat{\theta}}$	1.052	.44
$\hat{\beta}$	max	3.911	1.470
	mean	.830	.000
	min	-.966	-.961
	$S_{\hat{\beta}}$	1.000	.521
$\hat{\alpha}$	max	.514	
	mean	.321	
	min	.146	
	$S_{\hat{\alpha}}$	.083	

which approximates the true value of $-.5$. The observed discrepancy is due in part to the fact that the BICAL estimate $\bar{\hat{\theta}}_R = -.43$ involved some error. The standard deviation of the ability estimates becomes

$$S_{\hat{\theta}_T} = \frac{1.0}{.4595}(.44) = .958 \qquad [15]$$

which is quite close to the underlying value of unity.

The LOGIST results also were transformed to the θ_T metric via Equation 6 using the underlying values of $\bar{\alpha}_T$, $\bar{\beta}_T$, yielding

$$\bar{\hat{\theta}}_T = \frac{.321}{.270}(-.016)\left(\frac{1.052}{1.0}\right) + \left[.464 - \frac{.321}{.270}(.830)\left(\frac{1.052}{1.0}\right)\right] = -.574 \qquad [16]$$

Again, the value is close to the underlying value of $\bar{\theta}_T = -.5$. The standard deviation of the estimated ability scores becomes

$$S_{\hat{\theta}_T} = \frac{.321}{.270}(1.052) = 1.251. \qquad [17]$$

This value is a considerable departure from the underlying value and reflects the rather large error in the average value of $\bar{\hat{\alpha}}_N$ yielded by LOGIST.

Due to the anchoring procedures employed, neither the BICAL nor LOGIST programs are capable of recovering the absolute location of the ability score frequency distribution on the underlying ability scale. The best either can do is to recover the mean of the ability score frequency distribution relative to the average item difficulty. From Equation 6, for the BICAL results,

$$\left|\bar{\hat{\theta}}_T - \bar{\hat{\beta}}_T\right| = \left|\frac{1.0}{1.702(.270)}(-.43 - 0)\right| = .94 \qquad [18]$$

and for the LOGIST results,

$$\left|\bar{\hat{\theta}}_T - \bar{\hat{\beta}}_T\right| = \left|\frac{.321}{.270}(-.016 - .830)\frac{1.052}{1.0}\right| = 1.058 \qquad [19]$$

both of which are reasonably close to the true absolute difference of .964.

The results yielded by BICAL and LOGIST for this second data set show that despite the differences in anchoring procedures, both programs can recover a mean value of the ability score frequency distribution that transforms to the appropriate underlying value in the θ_T metric. BICAL did so despite the fact that the average value of α_T (.46 in logistic terms) was considerably less than unity. However, LOGIST did less well at recovering the variability of this distribution in the θ_T metric, illustrating the importance of obtaining an accurate estimate of the mean value of the α_i for use in the metric recovery process.

Data Set Involving Extreme Groups

The final example employed extreme groups and illustrates a more complex metric transformation situation in which the frequency distribution of the ability scores was involved. A third population of 1,100 simulated examinees, whose true ability scores were uniformly distributed over the θ_T metric from -2.5 to $+2.5$, was used. There were 100 examinees at each of 11 ability levels that were spaced .5 units apart. The frequency distribution of these ability scores had a mean of zero and a standard deviation of 1.581. The simulated test consisted of 20 equivalent items ($\alpha_T = .5$, $\beta_T = .0$) based upon a normal ogive model for the common ICC. Two subgroups were created by selecting examinees from the total group. The high group consisted of the 500 examinees whose ability scores were equal to or greater than $+.5$ yielding $\bar{\theta}_T = 1.5$ and $\sigma_{\theta_T} = .707$. The low group consisted of the 500 examinees whose scores were equal to or less than $-.5$ and having $\bar{\theta}_T = -1.5$ and $\sigma_{\theta_T} = .707$. The three data sets—total, high, and low—were each analyzed using BICAL and LOGIST. Table 3 contains the summary statistics for the parameter estimates yielded by the two computer programs. In each analysis, examinees who had

Table 3
Summary of Parameter Estimates Yielded by BICAL and LOGIST for the Total, High and Low Groups

	BICAL		LOGIST		
Group	$\hat{\beta}_i$	$\hat{\theta}_j$	$\hat{\alpha}_i$	$\hat{\beta}_i$	$\hat{\theta}_j$
TOTAL (N = 1,077)					
max	.070	2.79	.986	.036	2.11
mean	-.001	.02	.858	-.018	0.00
min	-.130	-2.79	.746	-.118	-2.14
S.D.	.047	1.24	.077	.034	1.00
HIGH (N = 491)					
max	.214	2.79	1.229	-.762	3.57
mean	-.000	1.23	.508	-1.910	0.00
min	-.178	-.80	.136	-5.233	-2.43
S.D.	.101	.48	.274	1.012	1.00
LOW (N = 486)					
max	.123	.59	.877	4.880	2.49
mean	.002	-1.18	.471	1.902	0.00
min	-.205	-2.79	.191	.969	-4.01
S.D.	.080	.49	.218	1.004	1.00

scores of 0 or 20 were removed, leaving 1,077, 491, and 486 examinees in the total, high, and low groups, respectively.

Under the Rasch model, the examinees are pooled according to their raw scores and there is a conditional distribution of underlying ability levels at each raw score. LOGIST, however, pools the examinees by their underlying ability levels and there is a conditional distribution of raw scores at each ability level. The major consequence of this is that the variance of the examinees' estimated abilities yielded by BICAL and LOGIST have different underlying values. The appropriate target values of the means and standard deviations for the total, high, and low groups, based on the reduced group sizes, are reported in Table 4. It should be noted that, as expected, the means agree across the two methods, but that the standard deviations differ considerably.

Total group. In the previous two data sets the variability of the ability score distribution was matched to the unit of measurement of the ability scale. As a result, the D_2/D_1 term of Equation 6 always had a value of unity. In the present data set the standard deviation of the ability scores was not matched to the unit of measurement, and it was necessary to take this into account. The total group results were examined first to assess the impact of the variance of the uniform distribution of ability upon the transformation process. The summary statistics yielded by LOGIST were $\bar{\hat{\theta}}_N = 0$, $S_{\hat{\theta}_N} = 1$, $\bar{\hat{\beta}}_N = .0182$, and $\bar{\hat{\alpha}}_N = .858$. The impact of the anchoring process can be seen in the value $\bar{\hat{\alpha}}_N$, which did not agree with the underlying value of .5. Under the LOGIST anchoring procedure $\sigma_{\theta_T} = 1.558$ (from Table 4) became $\sigma_{\theta_N} = 1$. The effect of this was to compress the ability scale and thus elevate the obtained value of $\bar{\hat{\alpha}}_N$. The value of $\bar{\hat{\alpha}}_T$ can be obtained from $\bar{\hat{\alpha}}_N$ via Equation 10 by solving for $\bar{\hat{\alpha}}_T$ and then

$$\bar{\hat{\alpha}}_T = \bar{\hat{\alpha}}_N \left(\frac{\sigma_{\theta_N}}{\sigma_{\theta_T}}\right) = .858 \left(\frac{1.0}{1.558}\right) = .551 \qquad [20]$$

Table 4
Summary Statistics of Ability Scores for Total, High and Low Groups of Table 3 by Examinee Organization

Group	Organized by θ_T Levels	Organized by Raw Score Levels
Total (N = 1077)		
$\bar{\theta}_T$	.011	.011
σ_{θ_T}	1.558	1.467
High (N = 491)		
$\bar{\theta}_T$	1.483	1.491
σ_{θ_T}	.701	.517
Low (N = 486)		
$\bar{\theta}_T$	-1.474	-1.474
σ_{θ_T}	.700	.539

If the underlying value of $\bar{\alpha}_N$ was used rather than $\bar{\hat{\alpha}}_N$, the transformation is

$$\bar{\alpha}_T = \bar{\alpha}_N \left(\frac{\sigma_{\theta_N}}{\sigma_{\theta_T}}\right) = .779 \left(\frac{1.0}{1.558}\right) = .5 \qquad [21]$$

Thus, the value of $\bar{\hat{\alpha}}_N$ yielded by LOGIST was an overestimate. In this data set, the unit of measurement yielded by LOGIST encompassed too many units of the underlying θ_T metric, and this fact will need to be taken into account when transforming the θ_N metric.

The summary statistics of the obtained ability estimates based upon the 1,077 examinees can be transformed to the θ_T metric via Equation 6, and the estimated mean is

$$\bar{\hat{\theta}}_T = \frac{.858}{.5}(0)\left(\frac{1}{1.558}\right) + \left[0 - \left(\frac{.858}{.5}\right)(-.0182)\left(\frac{1}{1.558}\right)\right] = .020. \qquad [22]$$

The standard deviation is

$$S_{\hat{\theta}_T} = \frac{.858}{.5}(1.0) = 1.716. \qquad [23]$$

The transformed mean is a reasonably close estimate of the underlying value of .011 given in Table 4. The standard deviation overestimates the underlying value of 1.558 by about 10%, which reflects the error in $\bar{\hat{\alpha}}_N$, since

$$1.716 \left(\frac{.5}{.551}\right) = 1.557. \qquad [24]$$

The summary statistics yielded by BICAL were $\bar{\hat{\theta}}_R = .02$, $S_{\hat{\theta}_R} = 1.24$, and $\bar{\hat{\beta}}_R = .001$. Substituting the appropriate values in Equation 6 yields a mean in the θ_T metric of

$$\bar{\hat{\theta}}_T = \frac{1.0}{.851}(.02)\frac{1.0}{1.0} + \left[0 - \frac{1.0}{.851}(.001)\frac{1.0}{1.0}\right] = .022 \qquad [25]$$

and the standard deviation is

$$S_{\hat{\theta}_T} = \frac{1.0}{.851}(1.24) = 1.457. \qquad [26]$$

Both values are quite close to the underlying values.

Extreme groups. The summary statistics of the ability estimates yielded by BICAL and LOGIST for the high and low groups were also reported in Table 3 and the target values in Table 4. For the high group, BICAL yielded $\bar{\hat{\theta}}_R = 1.23$, $S_{\hat{\theta}_R} = .48$, and $\hat{\beta} = 0$. In the θ_T metric the estimated mean ability becomes

$$\bar{\hat{\theta}}_T = \frac{1}{1.702(.5)}(1.23)\left(\frac{1}{1}\right) + \left[0 + \frac{1}{.851}(0)\right] = 1.445 \qquad [27]$$

which slightly underestimates the underlying value of 1.491. The standard deviation of the ability estimates becomes

$$S_{\hat{\theta}_T} = \frac{1}{1.702(.5)}(.48) = .564 \qquad [28]$$

in the θ_T metric and is an overestimate of the target value of .517 from Table 4. The summary statistics of the $\hat{\theta}_R$ yielded by BICAL for the low group were $\bar{\hat{\theta}}_R = -1.18$, $S_{\hat{\theta}_R} = .49$, and $\bar{\hat{\beta}}_R = .002$. The estimated mean in the θ_T metric is

$$\hat{\theta}_T = \frac{1}{1.702(.5)}(-1.18)\left(\frac{1}{1}\right) + \left[0 - \frac{1}{.851}(.002)\left(\frac{1}{1}\right)\right] = -1.384. \qquad [29]$$

The standard deviation is

$$S_{\hat{\theta}_T} = \frac{1}{.851}\ (.49) = .576. \qquad [30]$$

Both values are overestimates of their target values.

These extreme group results illustrate the importance of attending to the manner in which BICAL groups the examinees. While the extreme group means based upon raw score groupings and reported in Table 4 were similar to those based upon θ_T groupings, the standard deviations differed considerably under the two views of the underlying data. When the examinees were grouped by raw score, the σ_{θ_T} were roughly .53 for the extreme groups, which is considerably less than the .70 for examinees organized by θ_T values. However, the transformed values of $\sigma_{\hat{\theta}_R}$ indicated that BICAL was recovering its view of the underlying data reasonably well. This grouping effect also underlies the consistent underestimation of σ_{θ_T} observed in the first two data sets.

Due to the anchoring procedure used by LOGIST, transforming the extreme group results to the θ_T metric is rather involved. It will be useful to examine the impact of the anchoring procedures involved upon the value of $\bar{\alpha}_T$ before transforming the extreme group summary statistics to the θ_T metric. For the high group, the overall effect can be illustrated via Equation 10 as follows:

$$\bar{\alpha}_N = \bar{\alpha}_{NE}\ \left(\frac{\sigma_{\theta_{NE}}}{\sigma_{\theta_N}}\right) = .546\ \left(\frac{1}{.701}\right) = .779 \qquad [31]$$

but this value is drastically different from the underlying value of $\bar{\alpha}_T$. What has happened here is that the metric has been transformed to one in which the standard deviation of the ability scores is unity. It is important to note that the underlying θ_T metric has a unit of measurement of 1.0 but the standard deviation of the ability scores over this metric is not equal to 1.0. As a consequence, it is now necessary to transform this intermediate value of $\bar{\alpha}_N$ to that corresponding to the total group results and

$$\bar{\alpha}_T = \bar{\alpha}_N\ \left(\frac{\sigma_{\theta_N}}{\sigma_{\theta_T}}\right) = .779\ \left(\frac{1}{1.558}\right) = .500 \qquad [32]$$

is obtained. Thus, in the case of an extreme group a two-stage transformation process must be used.

For the high group of 491 examinees, LOGIST yielded $\hat{\theta}_{NE} = 0$, $S_{\hat{\theta}_{NE}} = 1.00$, $\bar{\beta}_{NE} = -1.91$, and $\bar{\alpha}_{NE} = .508$. The obtained average value of $\bar{\alpha}_{NE}$ was an underestimate of the expected value of .546. The first stage of the transformation of the obtained mean ability used the following values from Tables 3 and 4:

$$\theta_1 = \bar{\hat{\theta}}_N,\ D_1 = \sigma_{\theta_E} = .701,\ \bar{\alpha}_1 = \bar{\alpha}_N = .779,\ \bar{\beta}_1 = \bar{\beta}_T = 0 \qquad [33]$$

$$\theta_2 = \bar{\hat{\theta}}_{NE},\ D_2 = \sigma_{\theta_{NE}} = 1,\ \bar{\alpha}_2 = \bar{\hat{\alpha}}_{NE} = .508,\ \bar{\beta}_2 = \bar{\hat{\beta}}_{NE} = -1.91 \qquad [34]$$

Substituting these values in Equation 6 yields

$$\bar{\hat{\theta}}_N = \frac{.508}{.779}\ (0)\left(\frac{1}{.701}\right) + \left[0 + \frac{.508}{.779}\ (-1.91)\left(\frac{1}{.701}\right)\right] = 1.777 \qquad [35]$$

which is the mean expressed in terms of standard deviation units having a value of 1.0. The second stage of the transformation used the following values:

$$\theta_1 = \bar{\hat{\theta}}_T,\ D_1 = \sigma_{\theta_T} = 1.558,\ \bar{\alpha}_1 = \bar{\alpha}_T = .5,\ \bar{\beta}_1 = \bar{\beta}_T = 0 \qquad [36]$$

$$\theta_2 = \bar{\hat{\theta}}_N = 1.777,\ D_2 = S_{\hat{\theta}_N} = 1.0,\ \bar{\alpha}_2 = \bar{\hat{\alpha}}_N = .725,\ \bar{\beta}_2 = \bar{\beta}_N = 0 \qquad [37]$$

where .725 = .508/.701 and estimates .779. Substituting these values in Equation 6 yields

$$\bar{\hat{\theta}}_T = \frac{.725}{.5}(1.777)\left(\frac{1}{1.558}\right) + \left[0 + \frac{.725}{.5}(0)\left(\frac{1}{1.558}\right)\right] = 1.631 \qquad [38]$$

which is the mean of the high group expressed in the θ_T metric and the obtained value is an overestimate of the underlying value of 1.483. It should be noted that the estimates $\bar{\hat{\alpha}}_N$ and $\bar{\hat{\beta}}_N$ were used here rather than the values of $\bar{\alpha}_N$ and $\bar{\beta}_N$ and estimation errors have been compounded. The obtained standard deviation of the ability estimates can be transformed via Equation 10 as follows:

$$S_{\hat{\theta}_E} = \frac{.508}{.779}(1) = .652 \qquad [39]$$

which estimates .701 and then

$$S_{\hat{\theta}_T} = \frac{.725}{.5}(1) = 1.450 \qquad [40]$$

which estimates 1.558. For the low group of 486 examinees, LOGIST yielded $\bar{\hat{\theta}}_{NE} = 0.0$, $S_{\hat{\theta}_{NE}} = 1.00$, $\bar{\hat{\beta}}_{NE} = 1.092$, and $\bar{\hat{\alpha}}_{NE} = .471$. The first stage of the transformation of the observed mean ability was performed using Equation 6 and was

$$\bar{\hat{\theta}}_N = \frac{.471}{.779}(0)\left(\frac{1}{.6997}\right) + \left[0 - \frac{.471}{.779}(1.902)\left(\frac{1}{.6997}\right)\right] = -1.643 \qquad [41]$$

The second-stage value was

$$\bar{\hat{\theta}}_T = \frac{.673}{.5}(-1.643)\left(\frac{1}{1.558}\right) = -1.419 \qquad [42]$$

which estimates the underlying value of -1.474. The obtained standard deviation of ability can be transformed via Equation 10 as follows:

$$S_{\hat{\theta}_N} = \frac{.471}{.779}(1) = .605 \qquad [43]$$

which estimates .700 and

$$S_{\hat{\theta}_T} = \frac{.673}{.5}(1) = 1.346 \qquad [44]$$

which estimates 1.558.

Transforming ability estimates from θ_R to the θ_N metric. In practical situations the examinee scores on the θ_T metric are not known; hence, the transformation of interest is from the θ_R to θ_N metric or the reverse. For illustrative purposes, the summary statistics yielded by BICAL will be transformed to the LOGIST θ_N metric.

For the total group the mean ability becomes

$$\bar{\hat{\theta}}_N = \frac{1}{1.702(.858)}(.02)\left(\frac{1.0}{1.0}\right) + [-.0182-(.685)(-.001)] = .005 \qquad [45]$$

The standard deviation is

$$S_{\hat{\theta}_N} = \frac{1.0}{1.460}(1.24) = .849 \qquad [46]$$

which underestimates the desired value of 1.0. For the high group

$$\bar{\hat{\theta}}_{NE} = \frac{1}{1.702(.508)}(1.23)\left(\frac{1.0}{1.0}\right)+\left[-1.91 - \frac{1.0}{.865}(0)\left(\frac{1.0}{1.0}\right)\right]=.487 \quad [47]$$

$$S_{\hat{\theta}_{NE}} = \frac{1}{.865}(.48) = .555 \quad [48]$$

For the low group

$$\bar{\hat{\theta}}_{NE} = \frac{1.0}{1.702(.471)}(-1.18)\left(\frac{1}{1}\right)+\left[1.902 - \frac{1}{.802}(.002)\right] = -.428 \quad [49]$$

$$S_{\hat{\theta}_{NE}} = \frac{1}{.802}(.49) = .611 \quad [50]$$

In both groups the transformed values of $\bar{\hat{\theta}}_{RE}$ and $S_{\hat{\theta}RE}$ are poor estimates of the desired values of $\bar{\hat{\theta}}_{NE} = 0$ and $S_{\hat{\theta}NE} = 1$.

The transformation of the summary statistics of the ability estimates from the $\hat{\theta}_R$ to the $\hat{\theta}_N$ metric was not particularly successful. The major problem is that all the terms in the transformation equations are based upon estimates and the estimation errors are compounded. In particular, the value of $\bar{\alpha}$ yielded by LOGIST was in error by about 10%. In the case of the standard deviations, a contributing factor was the difference in organization of the examinees employed by the two programs.

Discussion

These results show that both the LOGIST and BICAL computer programs recovered a reasonably accurate linear function of the underlying θ_T metric for the data sets employed. That they do is not surprising, as both programs implement the same Birnbaum (1968) paradigm for the joint estimation of the item and examinee parameters. However, the two implementations differ with respect to the anchoring procedures used within the iterative estimation procedure. The anchoring procedure used by BICAL depends only upon the ICC parameters and the transformation of the unit of measurement in the θ_R metric to the θ_T metric depended only upon the ratio $1/(1.702\ \bar{\alpha}_T)$. When the value of $\bar{\alpha}_T$ was known, the transformed BICAL results were in good agreement with the underlying values, clearly indicating the role played by the average value of the item discrimination index in determining the unit of measurement. However, $\bar{\alpha}_T$ is not normally known, and an estimate must be obtained from an external source such as LOGIST. It should be noted that the discrimination estimates yielded by BICAL cannot be used as they are in the θ_R metric. The anchoring procedure used by LOGIST depends upon the frequency distribution of the examinees' estimated ability scores and sets $\bar{\hat{\theta}}_N = 0$ and $S_{\hat{\theta}_N} = 1$. The transformation of the θ_N metric to the θ_T metric depended upon the ratio $\bar{\alpha}_N\sigma_{\theta_N} / \bar{\alpha}_T\sigma_{\theta_T}$. The agreement of the transformed LOGIST results with the underlying θ_T values was a bit inconsistent. The basic problem was that the average value of the obtained estimates of the α_i differed slightly from the true value, and the discrepancy had a considerable impact upon the metric transformation process. Since sample sizes of 1,100 were used in the present study, it appears that LOGIST requires much larger sample sizes to obtain sufficiently precise estimates of α_i to be used in the metric transformation process.

The extreme group results yielded by the third data set exposed the importance of attending to the differences in the manner in which the examinees are grouped by the two computer programs. From Table 4 it was evident that the target value of S_{θ_T} for BICAL was less than the target value for LOGIST and markedly so for the extreme groups. Because of this, it appears that the difference in target values

of the ability score variances needs to be taken into account when comparing and interpreting results yielded by the two programs.

The transformation of the LOGIST results from the extreme group metric to the underlying metric revealed a rather subtle consequence of the anchoring procedure employed. It was found that a two-stage transformation process was necessary, whereas the same transformation of the BICAL results employed only a single stage.

The whole issue of metric transformation is encompassed by the latent trait/item response theory principle of invariance. If the item parameters are group invariant and the examinees' ability estimates are item invariant, then in theory transforming results from one metric to another can be readily accomplished. However, in applied situations, the several computer programs for estimating parameters employ different ICC models, anchoring procedures, and organizations of the examinees. Thus, if there is interest in comparing results across groups of examinees, across computer programs, or across combinations of these, it appears to be worthwhile attending to the metric issue using the procedures presented.

References

Baker, F. B. *GENIRV: A program to generate item response vectors*. Madison WI: University of Wisconsin, Laboratory of Experimental Design, 1978.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley, 1968.

Dinero, T. E., & Haertel, E. Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement,* 1977, *1,* 581–592.

Forsyth, R., Saisangjan, U., & Gilmer, J. Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement,* 1981, *5,* 175–186.

Guskey, T. R. Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement,* 1981, *5,* 187–201.

Hambleton, R. V., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement,* 1977, *14,* 75–96.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading MA: Addison-Wesley, 1968.

Loyd, B.H., & Hoover, H. D. Vertical equating using the Rasch model. *Journal of Educational Measurement,* 1980, *17,* 179–194.

Uhr, L. Pattern recognition computers as models for form perception. *Psychological Review,* 1963, *60,* 40–73.

Wood, R. Fitting the Rasch model—A heady tale. *British Journal of Mathematical and Statistical Psychology,* 1978, *31,* 27–32.

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

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton NJ: Educational Testing Service, 1968.

Wright, B. D., & Douglas, G. A. Best procedures for sample-free item analysis. *Applied Psychological Measurement,* 1977, *1,* 281–295.

Wright, B. D., & Mead, R. J. *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago IL: University of Chicago, Statistical Laboratory, Department of Education, 1977.

Wright, B. D., & Stone, M. H. *Best test design*. Chicago IL: MESA Press, 1979.

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

Author's Address

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