Estimating Item Characteristic Curves

Malcolm James Ree
Air Force Human Resources Laboratory

A simulation study of the effectiveness of four item characteristic curve estimation programs was conducted. Using the three-parameter logistic model, three groups of 2,000 simulated subjects were administered 80-item tests. These simulated test responses were then calibrated using the four programs. The estimated item parameters were compared to the known item parameters in four analyses for each program in all three data sets. It was concluded that the selection of an item calibration procedure should be dependent on the distribution of ability in the calibration sample, the later uses of the item parameters, and the computer resources available.

Increased interest in computer-driven adaptive testing, automated item banking, and automated test construction has made the estimation of the item characteristic curve (ICC) important. This curve describes the relationship between the ability of individuals and the probability of their answering a test question correctly. It is useful in estimating test scores, equating the scores of various tests, and scoring responses during adaptive testing. There are several methods for estimating ICCs within available computer programs. Selection and implementation of the appropriate program becomes a task for the practitioner. The objective of this study is to compare the merits of four available programs.

The Research Problem

In order to estimate an ICC, a conceptual model must be defined and item parameters must be estimated. The three-parameter logistic model of Birnbaum (Lord & Novick, 1968) is the most frequently used for relating item responses to persons' ability. The three parameters—*a*, *b*, and *c*—are item discrimination, item difficulty (or location), and probability of chance success (or lower asymptote), respectively.

The curve described by these parameters takes the shape of an (cumulative frequency) ogive or an "s" with the upper asymptote approaching a probability of 1.0 and, usually, the lower asymptote a probability greater than 0.0. The ogive describes the probability of obtaining a correct answer to an item as a monotonic increasing function of ability.

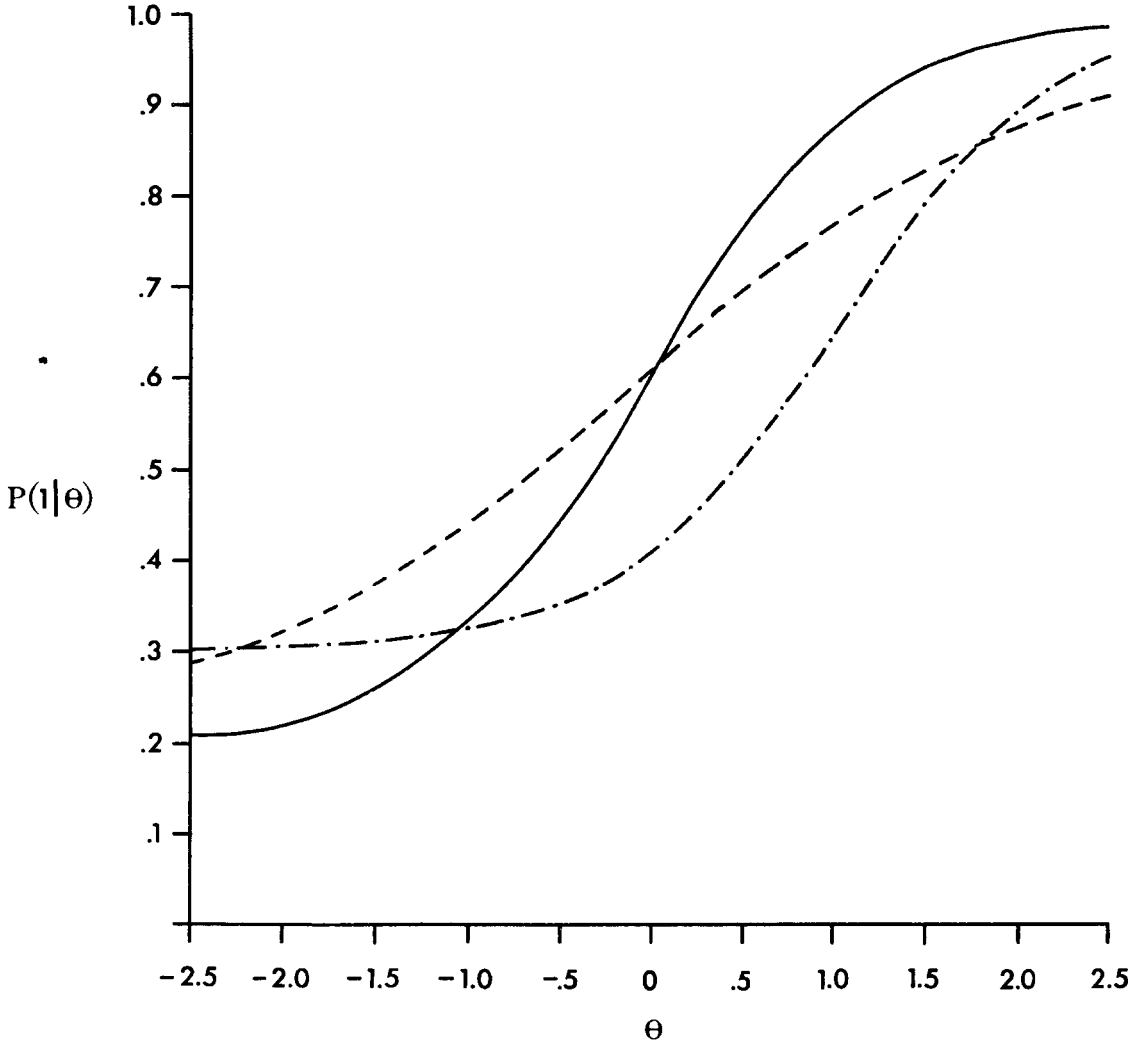
APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 3 Summer 1979 pp. 371-385

The item discrimination parameter (a) is a function of the slope of the ICC and generally ranges from .5 to about 2.5. A value of a equal to about 1.0 is typical of many test items, while a values below .5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are infrequently found. The item difficulty parameter (b) describes the point of inflection of the ICC and is usually scaled between −3.0 and +3.0, although the metric is arbitrary. The item guessing parameter (c) is the lower asymptote of the ICC and is generally interpreted as the probability of selecting the correct item-option by chance alone. Most test items have c parameters greater than 0.0 and less than or equal to .30.

Figure 1 shows three ICCs. The horizontal axis is scaled in units of ability (Θ) and the vertical axis is the probability of answering the item correctly. The solid curved line shows an ICC for an item

Figure 1

Item Characteristic Curves



of average difficulty with acceptable discrimination and the lower asymptote appropriate for a five-option multiple-choice item. The dashed line shows an item of identical difficulty, c value of .28, but with a lower a value. Note how the slope of the curve is less steep. The third curve, dot-dash line, shows an item with a c value of .30, an a parameter of 1.0, and a b parameter equal to 1.0. As the b parameter changes, the location of the inflection point of the curve is displaced along the horizontal axis.

In most cases the test constructor is faced with the task of estimating three parameters for the n items and one ability parameter (Θ) for every examinee (N) so that $N + 3n$ parameters must be estimated for each group of test items. For a group of 2,000 examinees taking 80 items 2,240 [2,000 + (3 × 80)] parameters must be estimated simultaneously. In an iterative procedure this estimation must be repeated several times, which leads to long computer runs with more precise estimates. Three of the four ICC estimation procedures evaluated in this study are iterative. The fourth is a monotonic increasing function of the biserial correlation between the item and the raw score.

Method

A simulation was run in order to have known values for the ability level (Θ) and for the item parameters. Three distributions of ability (Θ) with differing shapes were generated on which to test the procedures for ICC parameter estimation. Each Θ is equivalent to a "subject." The generated item parameters (a, b, and c) remained constant across the three distributions of ability (Θ).

Four methods of assessing the adequacy of the ICC estimation procedures were used. First, the estimated item parameters ($\hat{a}$, $\hat{b}$, and $\hat{c}$) were correlated with the known item parameters; second, the Θ estimated by using $\hat{a}$, $\hat{b}$, and $\hat{c}$ from each estimation procedure was correlated with the known Θ. Third, true scores and estimated true scores from the $\hat{a}$, $\hat{b}$, and $\hat{c}$ were compared (Lord, 1975). Finally, the test information curve was compared with estimates of the test information curve using the item parameters estimated in the three data sets.

Data Sets

Data set 1 (DS1). The distribution of Θ's for *DS1* was generated by dividing the interval between −2.5 and +2.5 into 2,000 equal intervals and assigning each resultant number to a value of Θ. This data set is similar to those sometimes produced for item analytic studies for tests such as the Armed Services Vocational Aptitude Battery (Jensen, Massey, & Valentine, 1976).

Data set 2 (DS2). The distribution of Θ's for *DS2* was generated by obtaining 3,000 cases from a unit normal random number generator. Two thousand Θ's were selected by administering a "test" and generating a sum of the number-correct score for the 3,000 based on ICC parameters of a 30-item subtest used in military selection and classification. A cutting score was set which would yield the upper two-thirds of the population. This method, rather than just cutting at a 33.3 percentile equivalent on the Θ distribution, was used to emulate actual selection practices which involve errors of measurement. The resultant distribution does not have a sharp truncation of Θ's but is asymmetric, with few scores below a specified level. *DS2* is similar to samples frequently available to organizations which must work with samples selected for inclusion in training or education.

Data set 3 (DS3). The distribution of Θ's for *DS3* was generated by accessing the unit normal random number generator for 2,000 numbers. Table 1 shows the means, standard deviations, and minimum and maximum Θ's for the three data sets.

Table 1
Descriptive Statistics for the Distribution of Theta for the Three Data Sets

Data Set	Mean	Standard Deviation	Minimum	Maximum	Skew	Kurtosis
1	-.0012	1.4437	-2.5000	2.4975	.0000	1.7991
2	.4957	.6998	- .5064	2.3791	.6359	2.7302
3	.0126	1.0191	-3.8445	3.6685	-.0050	3.1144

ICC Parameters

The distributions of ICC parameters were generated to simulate 80 five-option multiple-choice test questions. A normal distribution was specified for each ICC parameter. The means and standard deviations of these distributions were set to produce item parameters similar to those likely to be obtained in actual practice. Table 2 describes these distributions.

Table 2
Descriptive Statistics of the Generated ICC Parameters[a]

ICC Parameter	Mean	Standard Deviation	Minimum	Maximum
a	.9504	.2837	.6530	1.6136
b	.1635	.9286	-1.6530	1.9745
c	.2009	.0458	.0872	.3479

[a] These ICC parameters were used for all three data sets.

Generation of Item Responses

In order to generate a vector of item responses for each hypothetical testee the Θ values were used in Equation 1 to compute the likelihood of "passing" each item. The three-parameter logistic model is given by

$$P(\Theta)_j = c_i + (1 - c_i)\ (1 + e^{(-1.7a_i(\Theta - b_i))}) - 1 \qquad [1]$$

where $P(\Theta)_j$ is the probability of person j answering the test item correctly and a_i, b_i, and c_i are item parameters for item i.

Because Equation 1 yields a number, $P(\Theta)_j$, such that $0.0 < P(\Theta)_j < 1.0$, a number, X_j, was drawn from a uniform (rectangular) distribution ranging from 0.0 to 1.0 and compared to $P(\Theta)_j$. If X_j was larger than $P(\Theta)_j$, then an incorrect response was specified for the item; otherwise, a correct response was specified for the item. Thus, a hypothetical testee with $P(\Theta)_j = .90$ would get the item correct 9 in 10 times, and a vector of item responses was developed for each testee in each data set. These response vectors were then used to estimate *a*, *b*, and *c* by the four methods.

Estimation of ICC Parameters

The following four methods of ICC estimation were selected because of their wide availability to practitioners: ANCILLES, LOGIST, OGIVIA, and transformations to the item-test biserial correlation. All are three-parameter models.

ANCILLES and OGIVIA are described by Urry (1977, 1978); and LOGIST, by Wood, Wingersky, and Lord (1976). Transformations may be found in Lord and Novick (1968). These procedures were implemented on a UNIVAC 1108 and thoroughly checked out by processing the sample data set supplied by each of the authors of the programs. Default options for the programs were specified where possible and the logistic model was used throughout.

Analysis

Correlation of item parameters and Θ *estimates.* The first set of analyses consisted of correlating the ICC parameters with the estimated ICC parameters ($\hat{a}$, $\hat{b}$, and $\hat{c}$). The second set of analyses was of the correlation between Θ and $\hat{\Theta}$, where $\hat{\Theta}$ was computed using a maximum likelihood method and the various estimates of *a*, *b*, and *c* from the four procedures. These correlations were analyzed to determine how accurately Θ could be estimated from the $\hat{a}$'s, $\hat{b}$'s, and $\hat{c}$'s, as would be done in actual test administration with precalibrated items.

Maximum likelihood estimates (MLE) of Θ were computed using the likelihood function defined as

$$L(\Theta) = \prod (P(\Theta)^{u} Q(\Theta)^{1-u}) \qquad [2]$$

where $Q(\Theta) = 1 - P(\Theta)$, and u is 1 if the item was answered correctly and 0 if answered otherwise. The maximum of the distribution of likelihoods was found by the method derived by Jensema (1974). The use of this procedure is advantageous because it allows the estimation of Θ regardless of the sequence of item administration. Other methods, such as Bayesian estimation of Θ, are sequence dependent (see Sympson, 1977).

MLE is not sequence dependent but has the problems of possible failure to converge or reaching an asymptotically infinite estimate. Both of these problems can be palliated by arbitrarily placing a limit on the number of iterations and by placing an upper and lower limit on $\hat{\Theta}$. Maximum likelihood estimates of Θ were computed using the response vectors generated from Equation 1, each set of estimated item parameters, and the generated item parameters. The estimation of $\hat{\Theta}$ using the generated (*a*, *b*, and *c*) item parameters indicates the bias involved in the estimation of Θ alone. The correlation of Θ and the resultant $\hat{\Theta}$ is a measure of test reliability. No correlation of Θ and $\hat{\Theta}$ using any of the a's, $\hat{b}$'s, and $\hat{c}$'s should be expected to exceed the correlation of Θ and $\hat{\Theta}$ using the generated a's, b's, and c's.

True scores. The third set of analyses follows guidance proposed by Lord (1975) to eliminate most of the problems associated with estimating extreme values of Θ. These are termed true score (ξ) analyses. Because MLE procedures tend to exhibit bias on extreme cases, there may be a piling-up of high values at the minimum and maximum values allowed by the particular estimation routine. There are no empirical rules for setting either minimum or maximum values to be obtained in the MLE process; the limits set depend on judgment. In this study the values were set at −2.50 and +2.50. Other values might have yielded slightly different values. Estimation of true scores avoids these problems. Equation 3 defines true score:

$$\xi_j = \sum_{i=1}^{n} P_i(\Theta) \qquad [3]$$

where ξ_j is the true score,

n is the number of items, and

$P_i(\Theta)$ is the probability of a correct response for the item as in Equation 1. Similarly, the estimated true score is given by

$$\hat{\xi}_j = \sum_{i=1}^{n} P_i(\hat{\Theta}) \qquad [4]$$

where $P_i(\hat{\Theta})$ is computed from Equation 1 using a, b, and c.

Test information. The fourth set of analyses was comparisons of test information curves using the known a's, b's, and c's versus test information computed from $\hat{a}$, $\hat{b}$, and $\hat{c}$ from the four item parameter estimation techniques.

Item information is defined as

$$I_g(\Theta) = \left(\frac{\partial}{\partial\Theta} P_g(\Theta)\right)^2 \Big/ P_g(\Theta)\,(1 - P_g(\Theta)) \qquad [5]$$

where $P_g(\Theta)$ is estimated from Equation 1, and the numerator is the squared first derivative (i.e., the squared slope) of $P_g(\Theta)$ at a fixed value of Θ. Test information is the sum of the item information curves making up a test and is defined as

$$I(\Theta) = \sum_{i=1}^{n} I_g(\Theta) \qquad [6]$$

where $I_g(\Theta)$ is defined in Equation 5. Estimates of item information ($\hat{I}$) may be computed by substituting $\hat{a}$, $\hat{b}$, and $\hat{c}$ into Equation 1 and substituting that quantity into Equations 5 and 6.

It is useful to calculate item and test information curves in order to determine the precision of measurement of a test or an item. The height of the item or test information curve at any level of Θ may be thought of as being an ICC analog to classical measures of reliability. The higher the information curve the higher the information value and the higher the reliability of the item or test at that level of Θ.

Test information curves are frequently used to compare test characteristics (Brown & Weiss, 1977; McBride & Weiss, 1976; Vale & Weiss, 1977; Weiss, 1975) and to select items for administration during adaptive testing (Jensema, 1974; Ree, 1977). Because test and item information curves are computed using ICC parameters, errors of estimation of the parameters can cause test and item information curves to be incorrect. The item parameters are made comparable by placing them on common metric via a linear transformation of a and b. No such transformation of c is necessary.

Results

Item Parameters

Table 3 shows the results of the correlation of each of the estimated ICC parameters with the true ICC parameters. Note that the b parameter uniformly had the highest correlations and the c parameter had the lowest correlations with true values. The correlations observed in *DS2* show a decrement when compared with the other data sets. Specifically, the estimates of the a and c parameters were poor for all four estimation procedures.

Table 3
Correlations of ICC and Estimated ICC Parameters[a]

Procedure and Correlation	Data Set DS1	DS2[b]	DS3
ANCILLES			
$r(a.\hat{a})$	.873	.440	.836
$r(b.\hat{b})$	.960	.941	.968
$r(c.\hat{c})$	.409	.027	.325
LOGIST			
$r(a.\hat{a})$	.895	.565	.827
$r(b.\hat{b})$	.978	.447	.975
$r(c.\hat{c})$	.557	.233	.379
OGIVIA			
$r(a.\hat{a})$	.868	.556	.837
$r(b.\hat{b})$	.965	.923	.976
$r(c.\hat{c})$	.362	.000	.225
Transformation			
$r(a.\hat{a})$	.592	.323	.349
$r(b.\hat{b})$	.963	.917	.965
$r(c.\hat{c})$	c	c	c

[a] Following the rule of removing the items which LOGIST estimated to have b parameter values outside of the range of -3.0 to +3.0, three items were removed and the correlations recomputed for DS2 and are as follows: $r(a.\hat{a})$ = .571, $r(b.\hat{b})$ = .901, and $r(c.\hat{c})$ = .236. This observed increase was insufficient to warrant recomputation of the other analyses in like manner.

[b] Entries for ANCILLES and OGIVIA based on 75 and 64 items, respectively.

[c] Constant value of c = .20 precludes calculation of correlation.

The ANCILLFS procedure produced fairly consistent parameter estimates in *DS1* and *DS3* but did not produce the best estimate of any given parameter within a data set. LOGIST consistently produced the highest correlations of c and $\hat{c}$ and the highest correlations of a and $\hat{a}$ and b and $\hat{b}$ in *DS1*. The procedure produced the lowest correlation of b and $\hat{b}$ in *DS2*. The OGIVIA procedure produced the lowest intercorrelations between c and $\hat{c}$ of any of the procedures, while its estimation of a and b in *DS3* was superior to all others.

Ability Estimates

Table 4 shows the results of the Θ analyses. The column headed "Population" is the analysis using the generated item parameters and is presented as a reference set because no ICC calibration procedure can be expected to perform better than the true values.

The intercorrelation of number-correct score and Θ was exceeded by a small amount by the intercorrelations of $\hat{\Theta}$ and Θ only in *DS3*. In the other data sets the intercorrelation of number-correct score and Θ exceeded the intercorrelation of Θ and $\hat{\Theta}$, although this might be expected to change if the items were less well suited to the distribution of the ability of the subjects.

Only in *DS3* did the intercorrelation of Θ and $\hat{\Theta}$ for ANCILLES exceed the intercorrelation between number-correct score and Θ, and in no case did it outperform LOGIST or OGIVIA in this respect or in terms of the smallest average difference between Θ and $\hat{\Theta}$. The performance of LOGIST and OGIVIA was identical in *DS1* and *DS3* for correlation of Θ with $\hat{\Theta}$, but LOGIST consistently produced lower average differences between Θ and $\hat{\Theta}$.

Table 4
Descriptive Statistics for the Estimates of Θ Computed from the Generated and Estimated Item Parameters (N = 2,000)

Data Set and Statistic	S*	Estimation Method: Population	ANCILLES	LOGIST	OGIVIA	Transformation
Rectangular Data Set ($\mu\Theta$) = -.00125; $\Sigma(\Theta)$ = 1.4437						
Number of Items	80	80	80	80	80	80
$\bar{X}(\hat{\Theta})$	46.257	.0181	.0147	-.0133	.1004	-.0412
$\sigma(\hat{\Theta})$	19.629	1.4695	.0223	1.0163	.9087	.9038
$\bar{r}(\Theta.\hat{\Theta})$	.977	.980	.970	.974	.974	.955
$\Theta-\hat{\Theta}$		.0194	.0125	-.0121	.1016	-.0400
Skewed and Selected Data Set ($\mu\Theta$) = .49574; $\Sigma(\Theta)$ = .69989						
Number of Items	80	80	75	80	64	80
$\bar{X}(\hat{\Theta})$	52.565	.5028	-.0167	.0316	-.4219	.0199
$\sigma(\hat{\Theta})$	11.313	.7483	1.0147	1.0263	.9174	.9747
$\bar{r}(\Theta.\hat{\Theta})$	.939	.948	.935	.943	.937	.930
$\Theta-\hat{\Theta}$		-.0071	-.5123	-.4641	-.9176	-.4758
Normal Data Set ($\mu\Theta$) = -.01269; $\Sigma(\Theta)$ = 1.0191						
Number of Items	80	80	80	80	80	80
$\bar{X}(\hat{\Theta})$	45.587	.0096	-.0078	-.0073	.0706	-.0038
$\sigma(\hat{\Theta})$	14.615	1.0362	1.0020	1.0147	.9899	1.2313
$\bar{r}(\Theta.\hat{\Theta})$	.957	.966	.964	.965	.965	.961
$\Theta-\hat{\Theta}$		.0223	.0204	.0053	.0833	.0088

*Indicates number-correct score used as an estimate of Θ.

The transformations procedures worked best in *DS3*, producing a correlation of Θ with Θ higher than the correlation between number-correct score and Θ. In all other cases the correlation of Θ and $\hat{\Theta}$ for transformations was exceeded by the correlation of Θ and number-correct scores. Transformations did produce the smallest average difference between Θ and $\hat{\Theta}$ for *DS3*, but other procedures proved superior in *DS1* and *DS2*.

True Scores

Table 5 shows the means and standard deviations of ξ and $\hat{\xi}$, the average difference between them, and their intercorrelation.

All the procedures ranked ξ in much the same order, as evidenced by the correlation of ξ and $\hat{\xi}$, all exceeding .99. The greatest effect observed in this analysis was in average difference between ξ and $\hat{\xi}$. In all data sets the average difference measure for OGIVIA was less than 1.0, and only in *DS1* did another procedure, ANCILLES, produce a smaller average difference. Except in *DS3*, the transformations procedures showed the largest average difference between ξ and $\hat{\xi}$. In *DS1* and *DS2* the standard deviations of $\hat{\xi}$ were misestimated much more than in *DS3*, while the means of $\hat{\xi}$ were estimated well in all three data sets.

Information

Figures 2, 3, and 4 show the test information curve and estimates of the test information curve based upon $\hat{a}$, $\hat{b}$, and $\hat{c}$, estimated by the four methods in each of the data sets. In each of the figures,

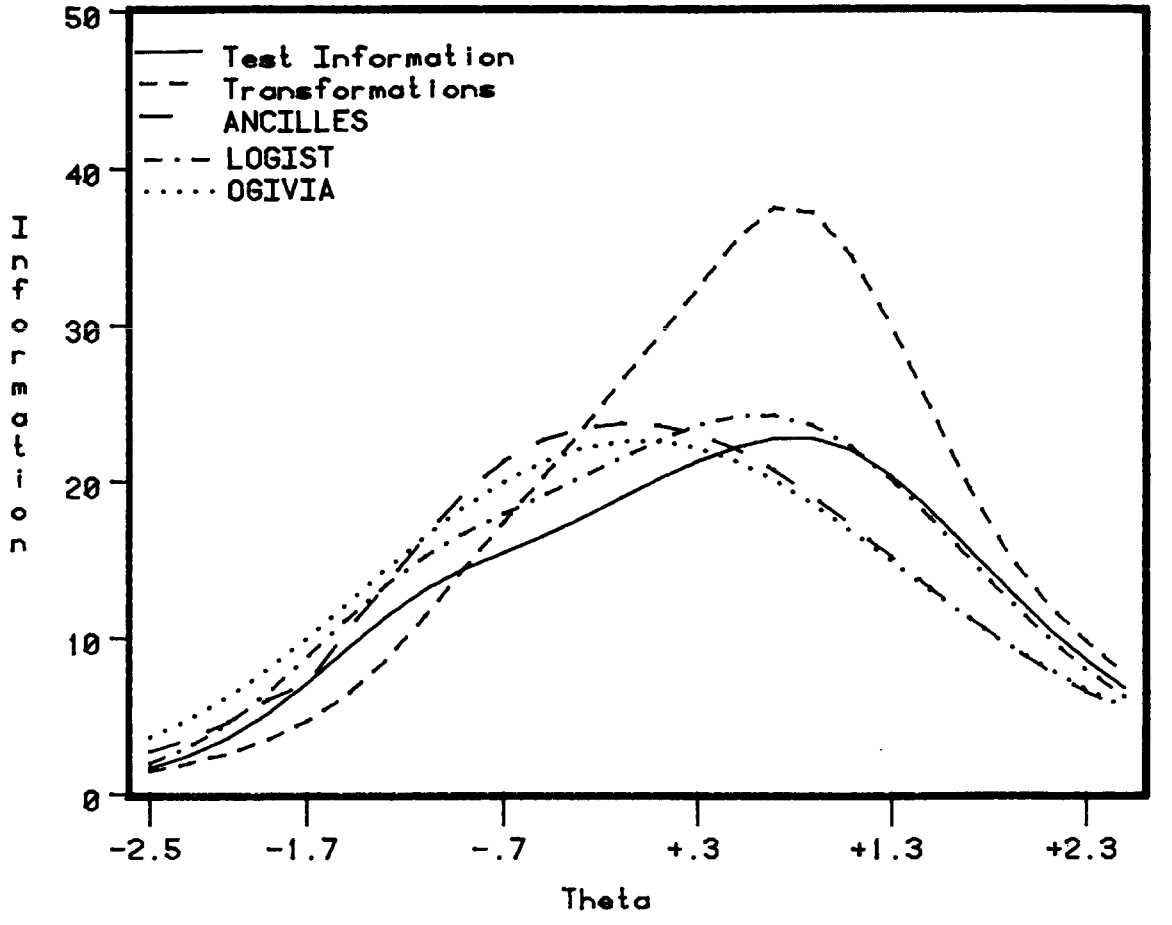
Table 5
Descriptive Statistics of ξ and $\hat{\xi}$, the Average Difference Between Them and Their Correlation

Procedure	$r(\xi.\hat{\xi})$	$\overline{X}(\hat{\xi})$	$\sigma(\hat{\xi})$	$\overline{(\xi - \hat{\xi})}$
Data Set 1 (μ_ξ = 46.009; Σ_ξ = 19.245)				
ANCILLES	.9927	46.444	24.927	.3444
LOGIST	.9960	47.205	23.424	1.1059
OGIVIA	.9945	45.210	25.091	- .8895
Transformation	.9910	47.589	24.352	1.4894
Data Set 2 (μ_ξ = 52.49; Σ_ξ = 10.592)				
ANCILLES*	.9995	54.63	7.783	5.3617
LOGIST	.9997	58.02	7.260	5.531
OGIVIA**	.9994	45.52	7.7895	- .4415
Transformation	.9999	58.04	8.028	5.550
Data Set 3 (μ_ξ = 45.326; Σ_ξ = 14.204)				
ANCILLES	.9998	45.900	14.325	.5737
LOGIST	.9999	46.085	14.112	.7591
OGIVIA	.9999	45.158	14.044	-.1680
Transformation	.9999	45.950	14.157	.6236

*75 items only
**64 items only

Figure 2

Test Information Curves, *DS1*



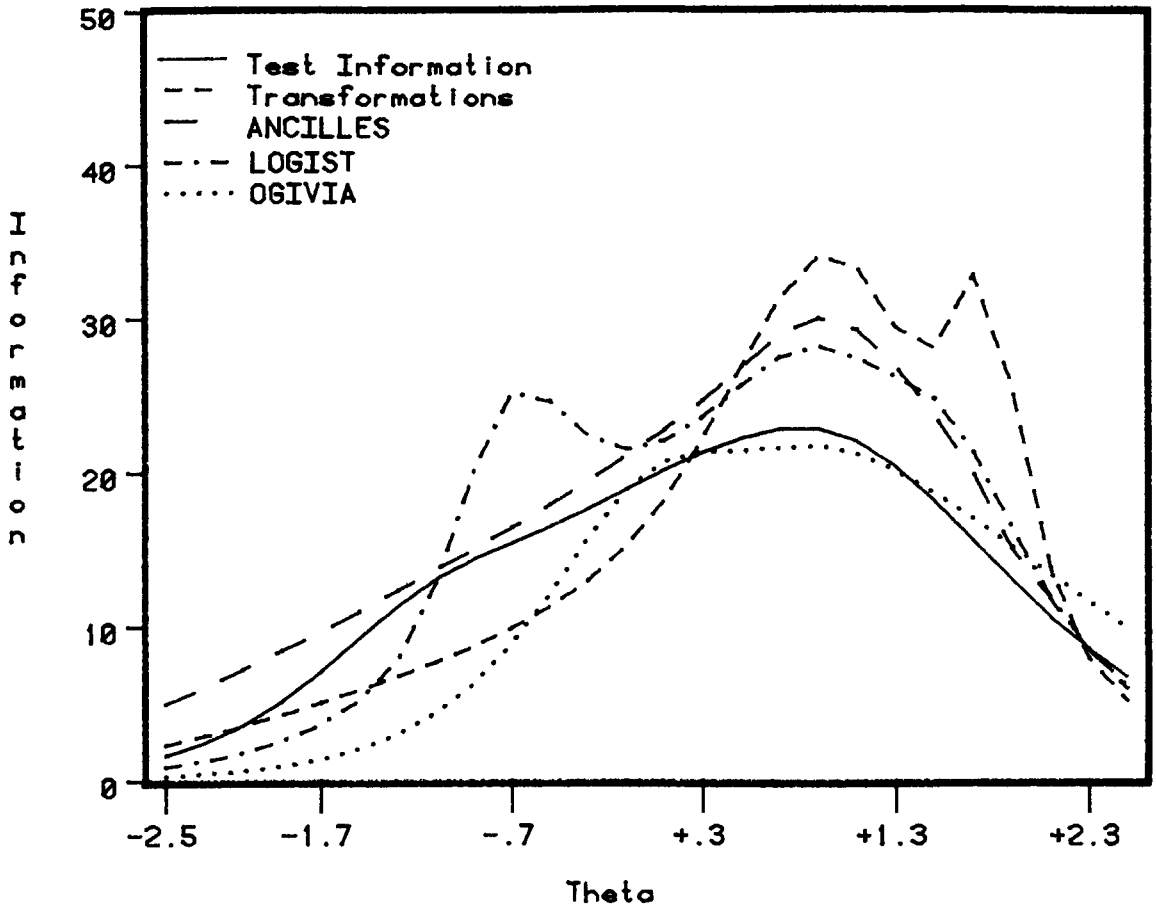
the line representing the information curve for $\hat{a}$, $\hat{b}$, and $\hat{c}$ calculated by the transformations procedures covers a larger area and is more peaked than the others. The other three procedures produced ICC parameters which estimated the information curve more accurately.

Table 6 presents the sum of squared deviations of true test information minus estimated test information as well as (1) the point on Θ where information reaches its maximum (Θ_g), (2) the correlation of I and $\hat{I}$, and (3) minimum and maximum values of $\hat{I}$ computed by each method in each of the data sets. All procedures in all the data sets except OGIVIA in *DS2* overestimated the total information. The smallest sum of squared differences between I and $\hat{I}$, the smallest difference between the estimated value of Θ at which I reached its maximum, and the true value of Θ at which I reached its maximum were found in *DS3*.

ANCILLES constantly overestimated the mean, minimum, and maximum information available. LOGIST overestimated the mean and maximum information in all data sets but underesti-

Figure 3

Test Information Curves, *DS2*



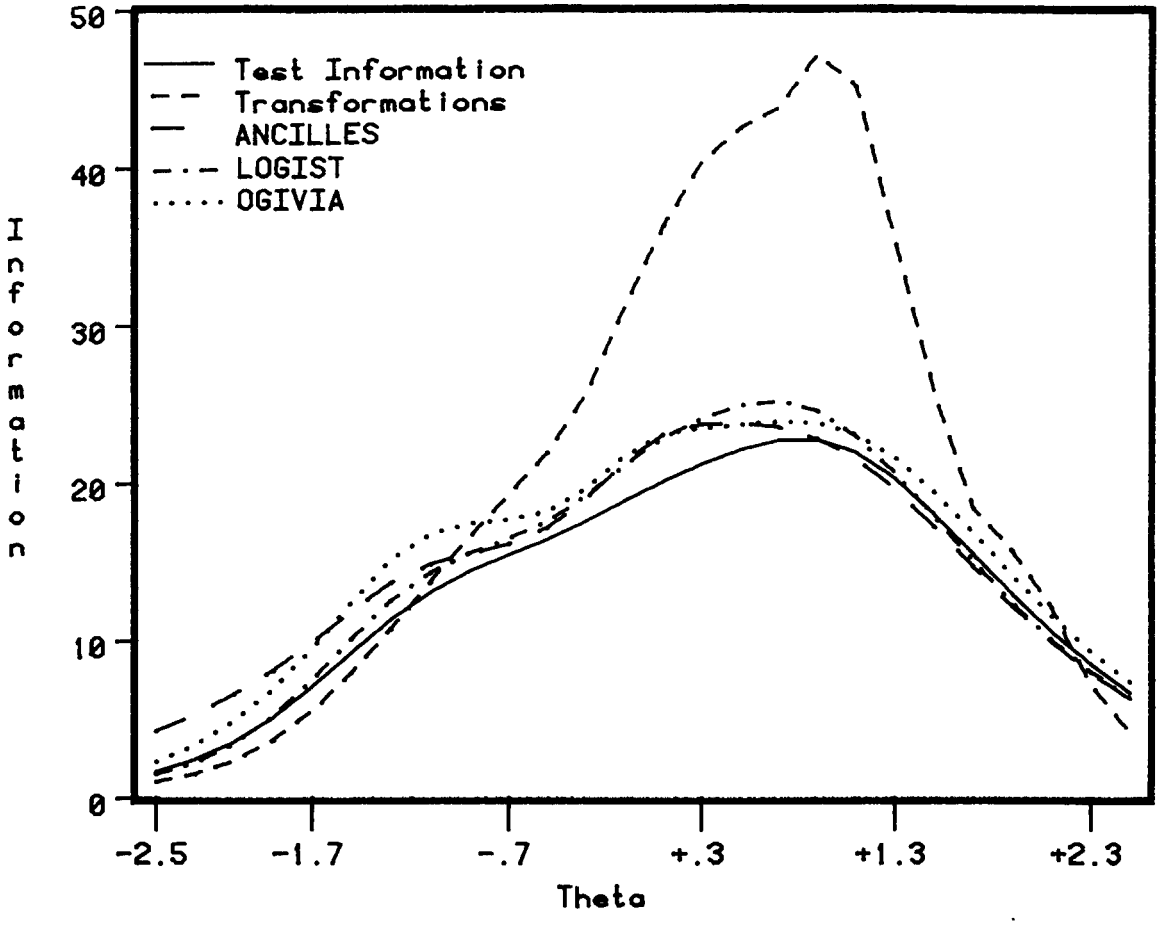
mated the minimum information in *DS2* and *DS3*. In *DS1* and *DS3* the LOGIST procedures produced the smallest sum of squared errors of the difference between information and estimated information. The intercorrelation between I and $\hat{I}$ was .95 or greater.

OGIVIA-estimated a, b, and c computed from *DS2* produced the only estimate of total mean, minimum, and maximum information lower than the actual value. In *DS1* and *DS2* the OGIVIA-based estimates of information correlated less than .900 with actual information. In no case did OGIVIA produce the lowest sum of squared differences between I and $\hat{I}$. In *DS2* and *DS3* the location of the maximum value of I was well estimated by OGIVIA.

Transformations-computed $\hat{a}$, $\hat{b}$, and $\hat{c}$ uniformly overestimated the amount of information and the maximum value of the information. Only in *DS1* was the correlation between I and $\hat{I}$ for the transformations procedures lowest; and in all cases, I and $\hat{I}$ was the highest for $\hat{a}$, $\hat{b}$, and $\hat{c}$ computed by the transformations procedures.

Figure 4

Test Information Curves, *DS3*



Discussion

The results clearly indicate that no one program functions best in all situations posed by the three data sets. Transformations performed poorly in most instances and is not recommended unless no other item calibration procedures are available.

In the rectangular data set (*DS1*) LOGIST produced results superior to the other procedures except in terms of the average differences between ξ and $\hat{\xi}$. The correlations of estimated item parameters and generated item parameters—Θ and $\hat{\Theta}$, I and $\hat{I}$, and ξ and $\hat{\xi}$—were higher for LOGIST than for any other procedure. LOGIST-estimated item parameters also most nearly reproduced the test information curve.

The results from the skewed and selected data set, *DS2,* call attention to a peculiarity exhibited by ANCILLES and OGIVIA. These two programs will not estimate parameters of some items under specific conditions. While this may seem a disadvantage, notice that $(\xi - \hat{\xi})$ for OGIVIA was the

Table 6

Information Analysis and Estimated Information Analyses Based on ICC Parameters

Data Set and Program	Total	Mean	Minimum	Maximum	Θ_g	$(I - \hat{I})$	$\Sigma(I - \hat{I})$	$r(I.\hat{I})$
Test Information	717.83	14.075	1.695	22.924	.800			
Estimated Test Information Based on ICC Parameters from DS1								
ANCILLES	736.31	14.437	2.757	23.708	-.100	-.362	650.04	.864
LOGIST	775.76	15.211	1.989	24.314	.600	-1.136	137.96	.986
OGIVIA	735.75	14.426	3.658	22.607	.000	-.351	621.29	.850
Transformation	930.77	18.250	1.477	37.708	.800	-4.175	2510.61	.971
Estimated Test Information Based on ICC Parameters from DS2								
ANCILLES*	871.24	17.083	5.016	29.954	.900	-3.008	694.15	.970
LOGIST	835.54	16.383	.954	28.096	-.600	-2.308	989.93	.958
OGIVIA**	613.24	12.024	.338	21.682	.900	2.051	854.63	.899
Transformation	806.66	15.817	2.360	34.105	1.00	-1.742	2361.61	.821
Estimated Test Information Based on ICC Parameters from DS3								
ANCILLES	777.81	15.251	4.280	23.906	.400	-1.1760	174.48	.976
LOGIST	762.35	14.948	1.539	25.300	.700	-.873	102.67	.994
OGIVIA	812.85	15.938	2.332	24.022	.800	-1.863	219.61	.991
Transformation	1070.70	20.993	1.046	47.565	1.00	-6.918	6416.10	.961

*75 items only

**64 items only

smallest in *DS2*. Note also that OGIVIA showed (see Table 4) a correlation of Θ with $\hat{\Theta}$ of .937 for 64 items compared to .943 for 80 items using LOGIST. This increase of .006 is very small for the addition of 16 items. LOGIST estimated item parameters for all the items, but inspection of the scatterplot of b versus $\hat{b}$ indicates several outliers which have the effect of substantially reducing the value of the correlation of b with $\hat{b}$. All the estimated test information curves computed from *DS2* estimates of the item parameters resembled the true test information curve very poorly.

The OGIVIA procedure was the most preferable for use in the normally distributed data set, *DS3*. The correlations of OGIVIA-estimated $\hat{a}$ and $\hat{b}$ with a and b were higher than for the other procedures; however, its correlation of c and $\hat{c}$ was less than either ANCILLES or LOGIST. The correlation of Θ with $\hat{\Theta}$ using OGIVIA was as high as LOGIST and higher than all others. The correlation of ξ with $\hat{\xi}$ for OGIVIA was the highest and simultaneously had the smallest average difference between ξ and $\hat{\xi}$. OGIVIA is built around assumptions of the normality of the distribution of Θ and performed very well when these conditions held true, as in *DS3*, or approximately held true, as in *DS2*. LOGIST estimates of the item parameters produced the highest correlation between I and $\hat{I}$ and the lowest sum of squared deviations of I minus $\hat{I}$, and thus the best estimated test information.

The decision as to which procedure to use must be based on a series of criteria. If all the items must be calibrated, then OGIVIA and ANCILLES may present problems in a situation similar to that represented by *DS2*. If wide-range samples like *DS1* and *DS3* are available or can be made available on the basis of ability measured on some other test, and the estimation of Θ is the goal, then calibration with LOGIST or OGIVIA is recommended. Clearly, if the examinees are available, a normal distribution of Θ leads to the best estimations of a, b, c, ξ, Θ, and I and is desirable. These data should then be calibrated using OGIVIA.

Cost is a final factor which should be considered. The transformations procedure was the quickest because, unlike the others, it is not iterative and its work can be accomplished in about 10 Fortran statements. The LOGIST procedure took the longest on the computer. It ran eight times longer than either ANCILLES or OGIVIA. Central Processor Unit (CPU) times on a UNIVAC 1108 with 262*K* words of memory for *DS3* were ANCILLES, 296 seconds; LOGIST, 2,061 seconds; OGIVIA, 180 seconds; and transformations, 38 seconds.

The data from this study, therefore, suggest that the choice of ICC parameter estimation techniques should be consistent with the later use of the estimates, the characteristics of the distribution of ability in the groups available for item administration, the necessity to calibrate all items, and the computer resources available.

References

Brown, J. M., & Weiss, D. *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A046062).

Jensema, C. An application of latent-trait mental test theory. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 29–48.

Jensen, H., Massey, I., & Valentine, L. *Armed Services Vocational Aptitude Battery Development: ASVAB Forms 5, 6, and 7* (AFHRL-TR-76-87). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division, December 1976.

Lord, F. *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Memorandum 75-33). Princeton, NJ: Educational Testing Service, 1975.

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

McBride, J., & Weiss, D. *Some properties of a Bayesian adaptive ability testing strategy* (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

Ree, M. Implementation of a model adaptive testing system at an Armed Forces Entrance and Examination Station. In D. J. Weiss (Ed.), *Proceed-*

ings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1977.

Sympson, J. Estimation of latent trait status in adaptive testing procedures. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing* (Research Report 77-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1977. (NTIS No. AD A038114)

Urry, V. *OGIVIA: Item parameter estimation program with normal ogive and logistic three-parameter model options*. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1977.

Urry, V. *ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options*. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1978.

Vale, C., & Weiss, D. *A comparison of information functions of multiple-choice and free-response vocabulary items* (Research Report 77-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1977.

Weiss, D. Adaptive testing research at Minnesota—overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U.S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

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

Acknowledgments

This reseach was conducted under Project 7719, Air Force Development of Selection, Assignment, Performance Evaluation, Retention, and Utilization Devices; Task 771915, Perceptual and Computer-Managed Measurement.

The author extends his appreciation to Vern Urry, United States Civil Service Commission, and to Frederick M. Lord, Educational Testing Service, for making their computer programs available and for their suggestions concerning the simulation and analyses. James R. McBride, Naval Personnel Research and Development Center; James B. Sympson, University of Minnesota; and Vincent Maurelli, Army Research Institute, provided much-appreciated assistance in the conduct of this study.

The views expressed herein are those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

Author's Address

Send requests for reprints or further information to Malcolm J. Ree, Personnel Research Division, Air Force Human Resources Laboratory, Brooks AFB, TX 78235.

Error Corrections

The errors indicated below appeared in the article

Estimating Item Characteristic Curve
by Malcolm James Ree
Volume 3, Number 3 (Summer 1979), pages 371-385.

Readers should remove this page and insert it in their copy of this article for future reference.

Page 374: Equation 1 should be

$$P(\theta)_j = c_i + (1 - c_i)[1 + e^{(-1.7a_i(\theta - b_i))}]^{-1} \qquad [1]$$

Page 376: The first line after Equation 4 should read
where $P_i(\hat{\Theta})$ is computed from Equation 1 using $\hat{a}$, $\hat{b}$, and $\hat{c}$.

Also on page 376: Equation 6 should be

$$I(\theta) = \sum_{i=1}^{n} I_g(\theta) \qquad [6]$$

Page 378: In Table 4 there were errors in labelling the last two rows of each section of the table. The row headings for the third and fourth rows of each of the three sections of Table 4 should be:

$r(\theta.\hat{\theta})$

$\bar{\theta} - \hat{\theta}$