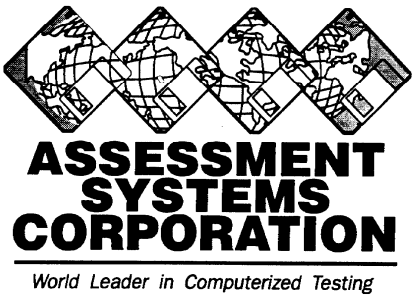
ASC Technical Report 95-1

An Updated Comparison of Microcomputer-Based Item Parameter Estimation Procedures Used with the 3-Parameter IRT Model

Michael E. Yoes

Assessment Systems Corporation

April 1995



Assessment Systems Corporation
2233 University Avenue, Suite 200
St. Paul, MN 55114
U.S.A.

Phone: (612) 647-9220 Fax: (612) 647-0412 E-Mail: asc@mr.net (Internet)

INTRODUCTION

There have been dramatic technical advances during the past twenty to thirty years in the field of educational and psychological testing. Paramount among these advances has been the ongoing development of item response theory (IRT). Although IRT holds a great deal of promise as a successor to more classical (i.e., true and error score) test theories, it has not been widely used by test practitioners.

The basis of IRT is the item response function (IRF; sometimes referred to as the item characteristic curve, ICC) which relates an examinee's trait/ability level (θ) to his/her probability of correctly answering the item. The IRF is a mathematical function that is defined by certain item parameters which, in practice, are never known but must be estimated from observed data. The process of item parameter estimation (sometimes referred to as item calibration) is one of the most difficult and important tasks in IRT and, unfortunately, one of the least understood.

Since item parameter estimation is so difficult, there is no single accepted method for accomplishing it. At least five theoretically distinct approaches have been taken with estimation in the 3-parameter IRT model. These item parameter estimation techniques include (1) approximation techniques, (2) minimum chi-square estimation techniques, (3) maximum-likelihood estimation techniques, (4) Bayesian estimation techniques, and (5) marginal maximum-likelihood estimation techniques. A detailed description of each technique is well beyond the scope of the present paper and the interested reader is referred to Baker (1987).

If IRT is to function in practical application as well as the theory predicts, accurate estimates of the item parameters are essential. Estimation of item parameters is, however, one of the major obstacles IRT poses for the practitioner. Due to the mathematical complexities involved, item parameter estimation is performed by computer software (programs) designed for that purpose. Until the early 1980s these estimation programs were available only for use on relatively "large" mainframe computer systems and were very costly to operate. This

posed a problem for many test practitioners who did not have the necessary computer access, programming support, and funding.

More recently, item parameter estimation programs have been developed for use on microcomputers. Microcomputer-based estimation promises to significantly reduce the cost, and hopefully the complexity, of parameter estimation and to place this capability onto the desktops of practitioners who otherwise would not be able to explore what IRT has to offer them.

To date little empirical research has been done to evaluate the accuracy and performance of these microcomputer-based item parameter estimation procedures (see for example: Yoes, 1990 and 1993; Vale and Gialluca, 1985 and 1988; Skaggs and Stevenson, 1986). The purpose of this paper is to update the Yoes (1990) study by presenting results on a new marginal maximum-likelihood estimation program (XCALIBRE) along with the results from the other estimation programs to summarize the effectiveness of the estimation procedures currently available for use with the 3-parameter IRT model.

METHOD

The present investigation was based on the methodology employed by Yoes (1990) in his comprehensive evaluation of two commercially available microcomputer-based item parameter estimation programs ASCAL (Assessment Systems Corporation, 1987) and BILOG (Scientific Software, 1986). The present paper extends the results of that study by incorporating evaluation of a new estimation program, XCALIBRE (Assessment Systems Corporation, 1995) under a subset of the conditions used in the original study. The following modifications of the design of the Yoes (1990) investigation were made for purposes of the present investigation: (1) the uniform distribution of θ condition was not used, (2) only test conditions 1 and 2 were evaluated, (3) the recovery of the averaged test information function was not evaluated, and (4) the recovery of the θ was not evaluated. Note that evaluation of item parameter estimation procedures are only possible using Monte Carlo simulation

techniques in which it is possible to specify and control the known parameters of both items and examinees, since those parameters are used to generate the data.

Generation of Item Response Data

All data sets were generated in accordance with the 3-parameter logistic IRT model in which the IRF is defined by the function:

$$P(u=1|\theta,a,b,c)=c+\frac{1-c}{1+e^{-Da(\theta-b)}} \qquad \text{(Equation 1)}$$

In this model the *a* parameter indexes an item's capacity to discriminate among differing levels of θ; the higher the *a* parameter, the more discriminating the item. The *a* parameter is commonly referred to as the item discrimination and is proportional to the slope of the IRF at its point of inflection.

The *b* parameter, referred to as the item difficulty, describes the item's location on the θ scale. Item difficulty is defined as the point on the θ scale at which the probability of a correct response is exactly half way between the upper and lower asymptotes of the IRF.

The *c* parameter is the probability that an examinee with a very low θ level would answer the item correctly through guessing or by some other means unrelated to the θ being measured. The *c* parameter can be referred to as the lower asymptote parameter. The scaling factor *D*, which is approximately equal to 1.7, maximizes the relation of the logistic ogive and the normal ogive models.

Item response data were generated for each simulated examinee by computing the probability of a correct response to the item in question given the examinee's "true" θ parameter and the "true" item parameters (using Equation 1). The probability of a correct response was then compared to a random number generated from a [0,1] uniform distribution using a random number generation procedure developed by Wickman and Hill (1982). If the probability of a correct response was less than the random number the item was coded as a correct response (1), otherwise the item was coded as incorrect (0).

Independent Variables

Yoes (1990) identified four major dimensions, or factors, which appear to be important in the evaluation of item parameter estimation procedures. Those factors are: (1) sample size, (2) test length, (3) distribution of θ, and (4) test characteristics. All of these factors, in one way or another, exercise an effect on the size, or content, of the observed (simulated) data matrix that is the basis from which all item parameter estimation procedures work.

Sample Size. Sample sizes used in the present study were N= 250, 500, 1000, and 2000. These were selected to span a range encompassing what might be considered to be a sample size too small for the 3-parameter model (N=250 condition) to that which might be used for development of an item bank (N=2000).

Test Length. Test Lengths were similarly chosen to reflect a broad range of test lengths representative of educational tests (i.e., n = 15, 25, 50, 75, and 100 items). Conventional wisdom or "rules of thumb" would classify a 15-item test as being too short for a 3-parameter item calibration.

Distribution of θ. In contrast to the previous (1990) study by Yoes where both normal and uniform distributions of θ were used only the Normal (0,1) distribution of θ was used in the present investigation. The decision to exclude the uniform distribution of θ condition was made based on issues of: (1) available time for conduct of the research, and (2) the previous finding that the distribution of θ had relatively little influence on the recovery of the item parameters. The decision to not include the uniform distribution of θ condition also avoided the necessity of addressing scaling issues as presented in the original paper (see Yoes, 1990 for a discussion).

Test Characteristics. Lord (1975) cautioned that simulation studies must include conditions that are representative of real data and testing situations. Test characteristics for the original study were selected to mirror item data that the author was working with at that time. Item discrimination (a) conditions were as follows: (1) Test 1 had moderate levels of item discrimination -- selected to mirror achievement data from an Introductory Psychology course in a large midwestern university (average a = 0.75, std. dev. = 0.1), and (2) Test 2

had high levels of item discrimination -- selected to mirror characteristics on a well developed instrument such as the General Science test from Armed Services Vocational Aptitude Battery (average a = 1.50, std. dev. = 0.2). In contrast to the original study, all item difficulties were distributed normally (0,1). Lower asymptote parameters were normally distributed with a mean of 0.25 and a standard deviation (SD) of 0.05 to reflect a 4-option multiple-choice exam. In combination, this produced two tests with characteristics as described in Table 1.

Table 1. Description of Test Characteristics used in generating response data.

Test Number	Item Parameter Distribution a	b	c
1	ND(0.75, .1)	ND(0, 1)	ND(.25, .05)
2	ND(1.50, .2)	ND(0, 1)	ND(.25, .05)

The four independent variables: (1) sample size, (2) test length, (3) distribution of θ, and (4) average level of item discrimination (i.e., test condition) were completely crossed, resulting in $4 X 5 X 1 X 2 = 40$ simulation data sets generated. Common data sets were used across the parameter estimation methods and the resulting item parameter estimates were compared to the known (generating) values to determine the effectiveness of each estimation procedure.

Estimation Procedures

Three commercially available (microcomputer-based) item parameter estimation programs, ASCAL (version 2.0; Assessment Systems Corporation, 1987) a maximum-likelihood procedure with Bayesian priors, BILOG (version 1.1; Mislevy and Bock, 1986) a marginal maximum-likelihood procedure, and XCALIBRE (version 1.0, Assessment Systems Corporation, 1995) a marginal maximum-likelihood procedure, were used to estimate the item parameters for each of the 40 data sets. As a relative performance comparison, the LOGIST 5 program (Wingersky, Barton, & Lord, 1982) a maximum-likelihood procedure operating on an

IBM mainframe computer was also used to estimate the item parameters for each of the data sets.

All programs were operated under their respective default options with the following exceptions: (1) BILOG was configured to characterize the distribution of θ empirically (rather than assuming a normal distribution of θ) and to allow the prior distributions on item parameters to be updated with each cycle of the estimation process (i.e., "floating" priors), (2) LOGIST estimates of θ were constrained in the interval [-4.0 to +4.0], (3) ASCAL was configured to allow a maximum of 25 "loops" in the estimation process, and (4) XCALIBRE was configured to allow a maximum of 25 "loops" in the estimation process and to allow the prior distributions on item parameters to "float" (as described for BILOG above).

Microcomputer estimation was carried out on IBM-compatible personal computers with math coprocessors. LOGIST runs were conducted on an IBM 3090 model 200 with Vector Facility. Computer time and support were generously provided by International Business Machines Corporation (IBM) through a research support program. All mainframe analyses were conducted at the IBM Los Angeles Scientific Center.

Evaluative Criteria

Individual Item Parameters. A number of criteria were used to evaluate the performance of each of the item parameter estimation programs. The first criterion was the recovery of each of the three individual IRT item parameters (***a***, ***b***, and ***c***) as indexed by (1) product-moment correlations ($\rho_{x\hat{x}}$) between the estimated and known parameter values, and (2) the root mean squared error (RMSE, the square root of the average squared difference between the true and estimated parameters). The RMSE index is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\pi_i - \hat{\pi}_i)}{n}} \qquad \text{(Equation 2)}$$

where π_i represents one of the "true" item parameters for item i (a, b, or c), $\hat{\pi}_i$ represents the estimate of the corresponding item parameter as produced by one of the four estimation methods, and n is the number of items in the data set (test length).

Item Response Function. Because errors in the individual item parameters can compensate for one another in fitting the observed data it was necessary to examine the recovery of the IRF as a whole. Recovery of the IRF was evaluated using a RMSE, the square root of the average squared distances between the true and estimated IRF at 201 points along the θ continuum (between $\theta = -2.50$ and $\theta = +2.50$ in increments of 0.025). For evaluating recovery of the IRF all items were used: items with item difficulties greater than 4.5 (in absolute value) were fixed at either -4.5 or +4.5, as appropriate. For evaluation of the IRF the RMSE index was defined as:

$$RMSE = \frac{1}{n}\sqrt{\frac{1}{201}\sum_{j=1}^{201}\left[P_i(\theta_j) - \hat{P}_i(\theta_j)\right]^2} \qquad \text{(Equation 3)}$$

Removal of Extreme Item Parameters

Any item difficulty estimates greater than 4.50 (in absolute value) were removed from the individual item parameter recovery analyses since there may be undue impact on the evaluation (particularly the RMSE criterion) due to a single item. This solution was devised to accommodate the fact that many of the LOGIST estimation runs resulted in 1 or 2 items that had "unreasonable" (i.e., very large) values. Table 2 shows the number of items removed for the item parameter recovery analyses; as can be seen in the table this problem pertained almost exclusively to the LOGIST program. The LOGIST program resulted in "extreme" difficulty estimates in 24 of the 40 estimation runs (60%). BILOG only resulted in a single "extreme" item difficulty in 40 estimation runs and the ASCAL and XCALIBRE programs always produced item difficulty estimates within "reasonable" limits.

Table 2. Numbers of Items Deleted from Item Parameter Recovery Analyses

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				Test 2			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	1	1	1	2	1	2	2	2
n2	0	1	0	0	1	1	0	1
n3	0	0	0	0	1	1	1	1
n4	0	1	0	0	2	0	1	0
n5	0	0	0	0	0	0	0	0
ASCAL								
n1	0	0	0	0	0	0	0	0
n2	0	0	0	0	0	0	0	0
n3	0	0	0	0	0	0	0	0
n4	0	0	0	0	0	0	0	0
n5	0	0	0	0	0	0	0	0
BILOG								
n1	0	0	0	0	0	0	0	0
n2	0	0	0	0	0	0	0	0
n3	0	0	0	0	0	0	0	0
n4	0	0	0	0	0	0	0	0
n5	0	0	0	0	0	1	0	0
XCALIBRE								
n1	0	0	0	0	0	0	0	0
n2	0	0	0	0	0	0	0	0
n3	0	0	0	0	0	0	0	0
n4	0	0	0	0	0	0	0	0
n5	0	0	0	0	0	0	0	0

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

Analysis of Variance

To assist in the evaluation of the estimation procedures it was desirable to consider the application of an analysis of variance (ANOVA) procedure. While ANOVA procedures have not typically been used in studies of this kind, it is more than likely attributable to the fact that the design of those studies did not permit the use of ANOVA. In the present design, the RMSE evaluation criterion for the IRF appeared to be a logical choice for the dependent measure for the ANOVA.

Examination of the distributions of the RMSE values, however, showed a significant skew indicating that the RMSE would require a logarithmic transformation prior to conducting the ANOVA procedure. The RMSE values were, therefore, transformed to log mean square error (LMSE) values using a base 10 logarithmic transformation of the squared RMSE values:

$$LMSE = \log_{10}\left(RMSE^2\right) \qquad \text{(Equation 4)}$$

For the recovery of the item response function, therefore, a full factorial ANOVA procedure was conducted with LMSE as the dependent variable and sample size, test length, estimation method, and test condition as the independent variables. The ANOVA was conducted using the ANOVA procedure from the SPSS-PC+ statistical program (version 5.0). Because there were no within-cell replications in the design, all 3-way and higher-order interaction effects were pooled to form a residual error term. For each main effect and 2-way interaction, an F test was conducted using the residual error term. The overall proportion of variance attributable to each effect (or interaction) was computed as η^2. Although RMSE values were transformed to LMSE for purposes of the ANOVA, mean values as reported (or graphed) are expressed on the original RMSE metric for purposes of interpretation.

RESULTS

Due to the scope of the study, general trends in the results will be noted along with any important findings. The interested reader will no doubt be able to glean many other findings from these data.

Recovery of Individual Item Parameters

Item Discrimination. Table 3 presents the product-moment correlations between estimated and known item discrimination values. As can be seen in the table, the correlations ranged considerably in value. In general, rank-order recovery of the true item discrimination values is not outstanding. Comparisons of the estimation procedures (not reported here due to length) show that item discrimination parameter estimates produced by each of the procedures were more highly correlated with each other than with the known values. Correlations between estimated and known item discrimination values tended to increase with increasing sample size and/or test length. On average, XCALIBRE had the highest correlations followed by BILOG, ASCAL and LOGIST.

Table 4 presents the RMSE information for recovery of the item discrimination parameter. The magnitudes of these values support the observation in Table 3 that the item discrimination parameter was not well estimated by any of the procedures, although the similarity among procedures is comforting. Note that the results of average (signed) bias computations (not reported here due to length) indicated that there was a tendency for all estimation procedures except for the XCALIBRE program to overestimate the item discrimination parameter.

Previous studies have shown a tendency for estimation procedures to occassionally allow the item discrimination parameter estimates to become unusually large. Table 5 presents the number of final item discrimination parameter estimates at or exceeding a value of 2.50 (the default maximum a value for LOGIST). The XCALIBRE and ASCAL programs appeared to perform better in constraining the values of the item discrimination parameters than the BILOG or LOGIST procedures.

Item Difficulty. The item difficulty parameters were well estimated by all four estimation procedures. The correlations in Table 6 show that rank-order recovery of the item difficulty parameters was very high.

Table 3. Recovery of the Item Discrimination Parameter (a) as Indexed by Product-Moment Correlations (*r*).

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				Test 2			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	.26	.02	.39	.67	.62	.72	.51	.91
n2	-.14	.61	.51	.45	.16	.40	.68	.67
n3	.30	.29	.50	.59	.37	.31	.62	.67
n4	.36	.33	.64	.69	.37	.47	.70	.69
n5	.25	.25	.44	.57	.40	.38	.45	.65
ASCAL								
n1	.47	.33	-.03	.51	.53	.45	.25	.73
n2	-.07	.57	.31	.44	.23	.12	.58	.43
n3	.28	.27	.44	.55	.36	.35	.55	.73
n4	.50	.45	.57	.69	.41	.48	.68	.71
n5	.34	.22	.46	.56	.53	.54	.63	.73
BILOG								
n1	.57	.55	.38	.66	.45	.43	.46	.69
n2	-.10	.66	.61	.54	.17	.04	.36	.62
n3	.42	.47	.57	.53	.22	.36	.50	.59
n4	.49	.44	.69	.58	.40	.51	.61	.58
n5	.40	.41	.52	.50	.49	.50	.59	.74
XCALIBRE								
n1	.59	.59	.42	.78	.51	.50	.49	.92
n2	-.14	.65	.60	.70	.38	.42	.59	.72
n3	.43	.47	.62	.70	.40	.55	.67	.78
n4	.54	.56	.73	.72	.44	.56	.73	.76
n5	.43	.45	.58	.69	.51	.53	.63	.75

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

Table 4. Recovery of the Item Discrimination Parameter (a) as Indexed by the Root Mean Square Error Criterion (RMSE).

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				Test 2			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	.72	.77	.34	.32	.66	.50	.57	.52
n2	.64	.37	.22	.31	.68	.45	.47	.29
n3	.45	.24	.21	.17	.55	.48	.26	.26
n4	.61	.37	.18	.13	.49	.38	.29	.21
n5	.48	.22	.22	.13	.57	.41	.32	.23
ASCAL								
n1	.69	.64	.75	1.08	.43	.49	.62	.66
n2	.54	.40	.34	.36	.43	.37	.40	.38
n3	.41	.24	.25	.19	.31	.31	.23	.19
n4	.38	.29	.19	.15	.27	.23	.21	.18
n5	.35	.25	.22	.15	.25	.23	.19	.17
BILOG								
n1	.19	.15	.14	.15	.61	.39	.35	.31
n2	.30	.12	.10	.13	.51	.38	.38	.21
n3	.20	.13	.13	.17	.46	.30	.25	.22
n4	.25	.23	.11	.15	.39	.26	.27	.23
n5	.20	.13	.15	.17	.48	.29	.22	.18
XCALIBRE								
n1	.09	.09	.11	.11	.20	.20	.21	.09
n2	.12	.08	.12	.09	.19	.22	.18	.18
n3	.10	.09	.09	.07	.20	.19	.17	.14
n4	.09	.08	.07	.07	.20	.17	.15	.13
n5	.09	.09	.08	.07	.19	.16	.15	.14

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

Table 5. Number of Item Discrimination Parameter Estimates at or exceeding 2.50 in Value

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				Test 2			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	2	1	0	0	7	3	4	4
n2	0	0	0	0	8	2	4	0
n3	0	0	0	0	6	6	1	2
n4	5	1	0	0	10	4	3	1
n5	2	0	0	0	21	6	2	2
ASCAL								
n1	0	0	0	2	0	0	0	2
n2	0	0	0	0	1	0	0	0
n3	0	0	0	0	0	0	0	0
n4	0	0	0	0	0	0	0	0
n5	0	0	0	0	0	0	0	0
BILOG								
n1	0	0	0	0	3	0	1	0
n2	0	0	0	0	1	1	1	0
n3	0	0	0	0	3	0	1	0
n4	0	0	0	0	5	1	1	0
n5	0	0	0	0	6	0	0	0
XCALIBRE								
n1	0	0	0	0	0	0	0	0
n2	0	0	0	0	0	0	0	0
n3	0	0	0	0	0	0	0	0
n4	0	0	0	0	0	0	0	0
n5	0	0	0	0	0	0	0	0

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

The error in estimation of item difficulties decreased with increases in either sample size or test length. BILOG and XCALIBRE tended to recover the item difficulty parameter better than did ASCAL or LOGIST in small sample conditions and/or short test lengths. There was a slight tendency for the high discrimination condition (Test 2) to recover the true item difficulty value with less error. In general, however, the results indicate that the programs are more similar than different for tests of 75 or 100 items and sample sizes of 1,000 or 2,000. Overall, XCALIBRE had the lowest RMSE, followed by BILOG, ASCAL, and LOGIST.

Lower Asymptote. Tables 8 and 9 report the correlational and RMSE evaluation (respectively) of recovery for the lower asymptote parameter (c). As is typically the case, the results suggest that the lower asymptote parameter may not be well estimated but there is also no clear distinction between the procedures in its estimation (although XCALIBRE and BILOG tended to result in lower RMSE values). As was the case for the item discrimination parameters, comparing estimates produced by each program showed a high degree of similarity (results not reported due to length; most correlations between estimates of any two of the procedures were > 0.70) between the programs in estimating the lower asymptote of the IRF. Rank-order recovery also was slightly affected by the average level of item discrimination -- the high a condition of Test 2 had a tendency toward better recovery of c as indicated by RMSE.

Recovery of the Item Response Function

In studies of this type it is perhaps most informative to look at relative performance of the estimation programs to focus attention on the recovery of the IRF as a whole (see Hulin, Lissak, and Drasgow, 1982). Table 10 shows the RMSE values for recovery of the entire IRF.

Table 6. Recovery of the Item Difficulty Parameter (*b*) as Indexed by Product-Moment Correlations (*r*).

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				Test 2			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	.97	.95	.98	.99	.97	.96	.99	.99
n2	.94	.96	.99	.99	.97	.99	.99	1.00
n3	.95	.95	.98	.99	.96	.99	1.00	.99
n4	.94	.96	.99	.99	.97	.99	.99	1.00
n5	.94	.97	.98	.99	.98	.98	.99	1.00
ASCAL								
n1	.98	.95	.97	.97	.98	.98	.98	.98
n2	.97	.97	.99	.99	.97	.99	.99	1.00
n3	.98	.97	.99	.99	.99	.99	.99	1.00
n4	.97	.98	.99	.99	.99	.99	1.00	1.00
n5	.97	.98	.98	.99	.99	.99	1.00	1.00
BILOG								
n1	.99	.98	.99	.99	.99	1.00	.99	.99
n2	.98	.98	.99	.99	.99	.99	1.00	1.00
n3	.98	.97	.99	.99	.99	.99	1.00	.99
n4	.97	.97	.99	.99	.98	.99	.99	.99
n5	.98	.98	.98	.98	.99	.99	1.00	1.00
XCALIBRE								
n1	.98	.98	.99	.95	.99	1.00	1.00	1.00
n2	.98	.98	.99	.99	.99	1.00	1.00	1.00
n3	.97	.98	.99	.99	.99	1.00	1.00	1.00
n4	.98	.98	.99	.99	.99	.99	1.00	1.00
n5	.98	.98	.99	.99	.99	.99	1.00	1.00

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

Table 7. Recovery of the Item Difficulty Parameter (*b*) as Indexed by the Root Mean Square Error Criterion (RMSE).

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				Test 2			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	.23	.29	.16	.16	.32	.72	.31	.38
n2	.35	.28	.20	.14	.31	.13	.18	.20
n3	.35	.34	.22	.15	.30	.18	.10	.12
n4	.35	.29	.19	.15	.26	.17	.12	.09
n5	.35	.29	.22	.16	.22	.21	.15	.12
ASCAL								
n1	.28	.33	.27	.40	.27	.20	.24	.23
n2	.24	.28	.15	.15	.26	.11	.11	.10
n3	.26	.27	.18	.14	.20	.14	.13	.10
n4	.26	.23	.17	.16	.17	.13	.10	.09
n5	.23	.24	.18	.15	.16	.12	.09	.09
BILOG								
n1	.18	.26	.19	.14	.19	.11	.15	.15
n2	.23	.29	.17	.15	.17	.11	.09	.08
n3	.25	.25	.16	.17	.21	.15	.11	.12
n4	.25	.25	.14	.17	.20	.14	.14	.12
n5	.23	.24	.18	.19	.17	.12	.09	.10
XCALIBRE								
n1	.22	.28	.19	.22	.23	.09	.11	.18
n2	.20	.37	.17	.14	.17	.12	.06	.10
n3	.25	.23	.13	.12	.16	.13	.10	.08
n4	.22	.21	.15	.16	.17	.11	.09	.06
n5	.21	.23	.14	.14	.15	.11	.08	.09

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

Table 8. Recovery of the Lower Asymptote Parameter (*c*) as Indexed by Product-Moment Correlations (*r*).

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				**Test 2**			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	.23	.20	-.20	-.09	.39	.60	.40	.54
n2	.13	.06	.61	.12	.18	.30	.68	.29
n3	.26	-.14	.07	.18	.28	.17	.45	.24
n4	.08	.14	.39	.29	.13	.19	.36	.53
n5	.06	.23	.06	.04	.18	.30	.34	.47
ASCAL								
n1	.43	.35	.06	.17	.36	.64	.34	.54
n2	.28	.14	.49	.37	.17	.34	.58	.37
n3	.27	.02	.34	.20	.43	.29	.41	.30
n4	.09	.21	.28	.32	.30	.31	.41	.54
n5	.21	.27	.20	.17	.28	.35	.47	.50
BILOG								
n1	.55	.33	-.08	.25	.13	.52	.21	.71
n2	.27	.18	.41	.38	.16	.33	.75	.57
n3	.24	-.01	.30	.13	.43	.23	.37	.27
n4	.10	.15	.31	.14	.36	.27	.46	.34
n5	.23	.32	.20	.20	.27	.36	.48	.51
XCALIBRE								
n1	.51	.38	.28	.35	.34	.60	.34	.67
n2	.32	.15	.27	.30	.16	.38	.80	.57
n3	.22	.09	.36	.30	.47	.27	.47	.37
n4	.10	.15	.25	.27	.32	.33	.45	.54
n5	.22	.36	.31	.29	.33	.43	.54	.59

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

Table 9. Recovery of the Lower Asymptote Parameter (c) as Indexed by the Root Mean Square Error Criterion (RMSE).

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				**Test 2**			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	.09	.11	.08	.06	.13	.09	.07	.07
n2	.09	.09	.06	.05	.10	.08	.07	.06
n3	.10	.09	.07	.05	.09	.09	.05	.06
n4	.11	.09	.05	.05	.11	.08	.05	.04
n5	.11	.08	.07	.05	.09	.07	.05	.04
ASCAL								
n1	.07	.07	.10	.14	.07	.08	.07	.08
n2	.05	.06	.05	.06	.06	.05	.06	.06
n3	.06	.05	.06	.05	.04	.05	.05	.05
n4	.06	.05	.05	.05	.05	.05	.04	.04
n5	.05	.05	.05	.05	.05	.05	.04	.04
BILOG								
n1	.03	.04	.04	.05	.04	.05	.05	.03
n2	.04	.05	.03	.03	.05	.04	.03	.03
n3	.04	.04	.04	.05	.03	.05	.04	.04
n4	.06	.05	.04	.05	.04	.04	.04	.04
n5	.04	.04	.05	.05	.04	.04	.04	.03
XCALIBRE								
n1	.03	.05	.03	.04	.05	.04	.04	.03
n2	.03	.07	.04	.04	.05	.03	.02	.03
n3	.03	.04	.04	.04	.03	.04	.04	.04
n4	.05	.04	.04	.05	.04	.04	.03	.03
n5	.04	.04	.03	.04	.03	.03	.03	.03

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

Overall, XCALIBRE resulted in the best recovery of the IRF as indexed by the RMSE criterion. For moderate discrimination condition (Tests 1), XCALIBRE parameter estimates resulted in a lower RMSE than those for BILOG in 12 out of 20 data sets for Test 1 and 16 out of 20 data sets for Test 2. The results from ASCAL were comparable to those obtained from LOGIST. For the high discrimination condition (Test 2) XCALIBRE and BILOG RMSE's were lower for short tests (n=15 and n=25 items) but as test length increased (n= 50, 75 and 100 items) results were comparable between all four programs.

Analysis of Variance

Results of the factorial analysis of variance (ANOVA) on the log mean square error (LMSE) for the IRF are presented in Table 11. What is of most interest in this ANOVA summary table are the η^2 values which indicate the size of the effect (proportion of total sum of squares accounted for). The independent variables and their 2-way interactions accounted for a proportion of 0.88 of the total variability in the LMSE values. As can be seen in Table 11 the largest effect was attributable to sample size (η^2 =.29). The mean values for the RMSE across sample sizes were: 250 examinees = .10, 500 examinees = .08, 1000 examinees = .06, 2000 examinees = .06. All four estimation procedures produced more accurate parameter estimates as sample size increased. Estimation method (program) accounted for the second largest proportion of the overall variability (η^2 =.20) with XCALIBRE and BILOG yielding noticeably lower RMSE values. The mean RMSE values for the estimation methods (across conditions) were: LOGIST = .10, ASCAL = .08, BILOG = .07, and XCALIBRE = .06. Test length also accounted for a sizable proportion of the overall variability (η^2 =.15) with decreasing RMSE values as test length increases. The mean RMSE values for test length conditions were: 15 items = .11, 25 items = .08, 50 items = .07, 75 items = .07, 100 items = .06. The sample size and test length effects provide empirical support of the statistical consistency of the estimation procedures. The effect of estimation procedure (program) appears to demonstrate the overall superiority of the marginal maximum-likelihood approach to estimating item parameters iin IRT. Both the BILOG and XCALIBRE

procedures recovered the overall IRF better than the other two procedures (ASCAL and LOGIST).

Table 10. Recovery of the Item Response Function (IRF) as Indexed by the Root Mean Squared Error Criterion (RMSE).

Estimation Procedure and Test Length	Type of Test and Sample Size							
	Test 1				**Test 2**			
	N1	N2	N3	N4	N1	N2	N3	N4
LOGIST								
n1	.15	.13	.10	.14	.18	.27	.18	.17
n2	.14	.10	.07	.06	.16	.09	.07	.09
n3	.12	.09	.06	.05	.10	.12	.06	.05
n4	.13	.09	.06	.05	.22	.08	.06	.04
n5	.12	.08	.07	.04	.12	.08	.06	.04
ASCAL								
n1	.15	.14	.14	.14	.12	.11	.12	.10
n2	.14	.10	.07	.08	.12	.07	.07	.07
n3	.11	.08	.06	.05	.09	.09	.06	.04
n4	.11	.08	.06	.05	.09	.07	.05	.04
n5	.10	.07	.06	.04	.09	.07	.05	.04
BILOG								
n1	.08	.07	.06	.06	.10	.07	.08	.05
n2	.10	.07	.04	.05	.10	.06	.05	.05
n3	.09	.07	.05	.05	.09	.08	.05	.05
n4	.10	.07	.05	.05	.10	.07	.06	.06
n5	.08	.06	.06	.05	.09	.07	.05	.05
XCALIBRE								
n1	.07	.07	.06	.07	.09	.07	.05	.06
n2	.08	.08	.06	.05	.08	.06	.04	.04
n3	.08	.06	.04	.03	.07	.07	.05	.04
n4	.08	.06	.05	.04	.08	.06	.04	.03
n5	.07	.06	.04	.03	.08	.06	.04	.04

NOTES:

Sample Sizes
N1 = 250
N2 = 500
N3 = 1,000
N4 = 2,000

Test Lengths
n1 = 15 items
n2 = 25 items
n3 = 50 items
n4 = 75 items
n5 = 100 items

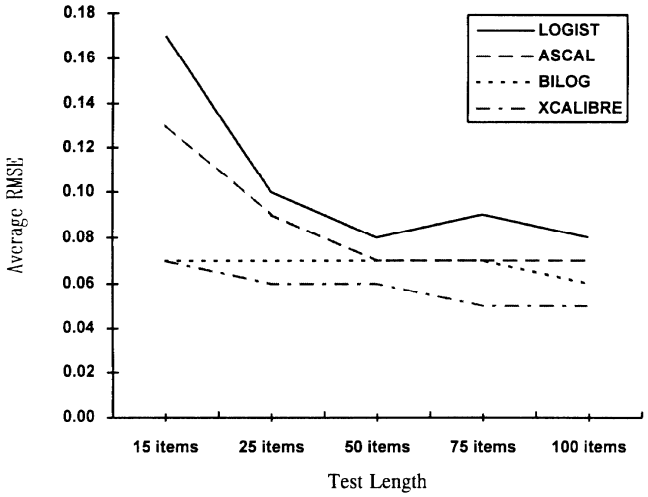
Table 11. Factorial Analysis of Variance (ANOVA) of the Log Mean Square Error (LMSE) criterion for the Item Response Function (IRF)

Source of Variation	Sum of Squares	df	Mean Square	F	Signif	η^2
Main Effects						
Sample Size	5.959	3	1.986	83.945	.000	0.29
Test Length	3.134	4	.783	33.112	.000	0.15
Estimation Method	4.170	3	1.390	58.746	.000	0.20
Test Condition	.064	1	.064	2.684	.104	0.00
2-way Interactions						
Sample Size X Test Length	1.292	12	.108	4.552	.000	0.06
Sample Size X Estimation Method	.221	9	.025	1.039	.414	0.01
Sample Size X Test Condition	.016	3	.005	.232	.874	0.00
Test Length X Estimation Method	1.755	12	.146	6.180	.000	0.09
Test Length X Test Condition	.124	4	.031	1.313	.270	0.01
Estimation Method X Test Condition	.143	3	.048	2.012	.117	0.01
Explained	17.926	54	.332	14.030	.000	0.88
Residual	2.484	105	.024			
Total	20.411	159	.128			

Notes: All three-way and higher interaction effects were pooled to form the residual term.
Of the 720 total cases, 21 cases contained missing data.

There were two sizable (i.e., $\eta^2 \geq .05$) 2-way interaction effects in the ANOVA. The largest interaction effect ($\eta^2 = .09$) was between test length and estimation method. This interaction is graphed in Figure 1 and again demonstrates the superiority of the estimation procedures based on marginal maximum-likelihood (BILOG and XCALIBRE). BILOG and XCALIBRE yielded RMSE values that were relatively constant across test length, thus giving empirical verification to the theoretical advantages of marginal maximum-likelihood with short tests; XCALIBRE showed a tendency toward reduced RMSE values as test length increased. Both LOGIST and ASCAL demonstrated a trend of decreasing RMSE values as test length increased but very relatively high levels of RMSE (in contrast to BILOG and XCALIBRE) for tests of 15 and 25 items.

Figure 1. ANOVA of LMSE for the IRF
2-Way Interaction between Test Length and Estimation Method $(\eta^2 = .09)$



The only remaining sizable interaction effect was between sample size and test length. The RMSE mean values for this interaction are presented in Figure 2.

Figure 2. ANOVA of LMSE for the IRF
2-Way Interaction between Sample Size and Test Length $(\eta^2 = .06)$

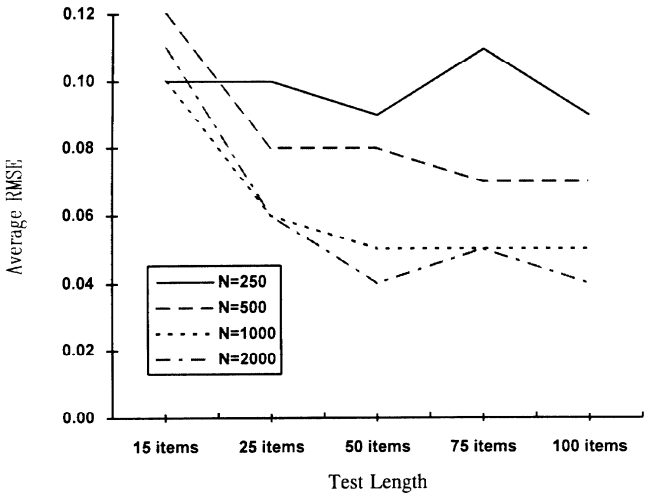


Figure 2 shows that with very small samples ($N = 250$) the lengthening of the test did not result in an overall reduction in the average RMSE across procedures. With a more moderate sample size (N=500) the effect of test length began to be demonstrated, but without as much reduction as was found in the N=1000 and N=2000 conditions. Note, however, that this effect is across all estimation procedures.

DISCUSSION

Although measures of recovery of the individual item parameters can help to shed light on the areas in which a particular estimation procedure/program may be faltering (e.g., overestimation of the item discrimination parameter by all of the programs *except* XCALIBRE), the recovery of the IRF as a whole appears to provide the best indicator of program performance. Across data sets, the marginal maximum-likelihood procedures (XCALIBRE and BILOG) resulted in improved recovery of the IRF, particularly in conditions where theoretically expected -- i.e., short test lengths and small sample sizes. As sample size increased to 500 or more, however, and test length increased to 50 items or more, the differences between the programs become less pronounced. However, Figure 1 shows that XCALIBRE produced the lowest mean RMSE at all test lengths above 15 items, and equaled that of BILOG at 15 items.

Overall the marginal maximum-likelihood estimation procedures (BILOG and XCALIBRE) would appear to be the best overall choice for an IRT parameter estimation program. This is particularly true when working with data sets resulting from short tests and/or small samples. Results from the present study demonstrate that the new XCALIBRE procedure is a viable alternative to BILOG.

Since the Yoes (1990) investigation was undertaken, new versions of both ASCAL and BILOG have been released. Updated versions of both ASCAL (version 3.2) and BILOG (version 3.04) were used on a subset of these same data sets. Based on this sampling, it did not appear that the results of this study would be affected by re-running the analyses on the updated programs.

REFERENCES

Assessment Systems Corporation (1987). *MICROCAT: A computer program for computerized adaptive testing*. St. Paul, MN: Author.

Assessment Systems Corporation (1989). *User's manual for the ASCAL 2- and 3-parameter IRT calibration program*. St. Paul, MN: Author.

Assessment Systems Corporation (1995). *User's manual for the XCALIBRE marginal maximum-likelihood estimation program*. St. Paul, MN: Author.

Baker, F.B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*, 111-141.

Hambleton, R.K., and Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hulin, C.L., Lissak, R.I., and Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*, 249-260.

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

Mislevy, R.J., and Bock, R.D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Skaggs, G., and Stevenson, J. (1986). *A comparison of ASCAL and LOGIST parameter estimation programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Vale, C.D., and Gialluca, K.A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters* (ONR-85-4). St. Paul: Assessment Systems Corporation.

Vale, C.D., and Gialluca, K.A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement*, *12*, 53-67.

Wickman, B.A., and Hill, I.D. (1982). An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society*, *31*, 188-190.

Wingersky, M.S., Barton, M.A., and Lord, F.M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.

Yoes, M.E. (1990). *A comparison of microcomputer-based item parameter estimation procedures used with the 3-parameter IRT model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Yoes, M.E. (1993). *A comparison of the effectiveness of item parameter estimation techniques used with the 3-parameter logistic item response theory model*. Unpublished doctoral dissertation, University of Minnesota.