Open access • Journal Article • DOI:10.1177/014662168200600301

Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study — Source link

Charles L. Hulin, Robin I. Lissak, Fritz Drasgow

Institutions: University of Illinois at Urbana–Champaign

Published on: 01 Jun 1982 - Applied Psychological Measurement (SAGE Publications)

Topics: Item response theory, Differential item functioning, Sample size determination and Mean squared error

Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study

Charles L. Hulin, Robin I. Lissak, and Fritz Drasgow
University of Illinois

This monte carlo study assessed the accuracy of simultaneous estimation of item and person parameters in item response theory. Item responses were simulated using the two- and three-parameter logistic models. Samples of 200, 500, 1,000, and 2,000 simulated examinees and tests of 15, 30, and 60 items were generated. Item and person parameters were then estimated using the appropriate model. The root mean squared error between recovered and actual item characteristic curves served as the principal measure of estimation accuracy for items. The accuracy of estimates of ability was assessed by both correlation and root mean squared error. The results indicate that minimum sample sizes and tests lengths depend upon the response model and the purposes of an investigation. With item responses generated by the two-parameter model, tests of 30 items and samples of 500 appear adequate for some purposes. Estimates of ability and item parameters were less accurate in small sample sizes when item responses were generated by the three-parameter logistic model. Here samples of 1,000 examinees with tests of 60 items seem to be required for highly accurate estimation. Tradeoffs between sample size and test length are apparent, however.

An important problem encountered in applications of item response theory (IRT) is estimation of person and item parameters. In most practical applications, both person and item parameters must be estimated simultaneously. The method of maximum likelihood is one procedure that has been used to estimate parameters of IRT models. Unfortunately, theorems that describe the usual properties of maximum likelihood estimates are not necessarily true when *both* person and item parameters must be estimated. One important property of an estimator is *consistency;* as sample size becomes large, a consistent estimator of a parameter converges to the parameter.[1] Under usual circumstances, maximum likelihood estimates are consistent (Kendall & Stuart, 1979), but a general proof of consistency for maximum likelihood estimates of IRT parameters has not been developed.

In this paper the properties of the LOGIST computer program (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976) are examined. LOGIST uses the method of maximum likelihood whenever feasible to estimate person and item parameters. Estimation for the two- and three-parameter logistic models is studied here because these models are widely used.

Lord (1968) has suggested that samples of $N > 1,000$ examinees and $n > 50$ items are needed for adequate estimation of one of the three item parameters of the three-parameter logistic model (item discrimination). Evidence

[1] More formally, a consistent estimator of a parameter converges in probability or converges stochastically to the parameter.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 3, Summer 1982, pp. 249–260

0146-6216/82/030249-12$1.60

supporting Lord's conjecture was provided by Swaminathan and Gifford (1979), who found that the item discrimination parameter was estimated poorly if $N = 50$ or 200 and $n = 10$, 15, or 20. This would imply that applications of IRT to attitude measurement problems are virtually impossible: Time requirements would be prohibitive if 50 or more items are required to measure a single attitude. Furthermore, the range of practical applications in ability and aptitude testing would also be greatly reduced.

Ree and Jensen (1980) have also found sample size requirements for item parameter estimation to be substantial for the three-parameter logistic model. They stated that "a stable and accurate estimate of the *a* and *b* parameters requires large numbers of subjects over a broad range of ability" (p. 227). Using simulated tests of length $n = 80$ items, they found estimation errors of the *a* (discrimination) parameter to be large in samples of less than $N = 1{,}000$ examinees. Lord (1980a) examined the (improper) use of the Rasch model for item responses generated by the two-parameter logistic model. He found that Rasch model estimates of the person parameter were superior to two-parameter logistic ability estimates when the testing norming sample was sufficiently small.

The methods used in this study for evaluating estimates of item parameters differ from those previously used. Lord (1975), Swaminathan and Gifford (1979), and Ree and Jensen (1980) all examined recovery of *item parameters* for the three-parameter logistic model. The present study examines *recovery of the item characteristic curve* (ICC) for the *two-* and three-parameter logistic models.

An analogy with multiple regression is apparent. Studying the recovery of item parameters in IRT corresponds to studying the recovery of regression equation coefficients (i.e., beta weights); the examination of recovery of the ICC corresponds to investigating the mean squared error of prediction in multiple regression. Interest in studying the recovery of the ICC rather than item parameters results from the analogy with multiple regression. First, note that in most applications of IRT, the main interest lies in the ICC; item parameters are only a convenient means for summarizing the ICC. Similarly, in applications of multiple regression, interest frequently lies in the predicted criterion scores. Here, regression equation coefficients conveniently summarize the regression hyperplane.

In multiple regression it has been found that large differences in regression coefficients have little influence on the parts of the regression hyperplane that are used to predict most criterion scores (Dorans & Drasgow, 1978; Wainer, 1976). Large estimation errors of IRT parameters may have only small influences on the parts of the ICC that are relevant to particular application. Linn, Levine, Hastings, and Wardrop (1981) provide an example of two hypothetical ICCs, one with $a = 1.8$, $b = 3.5$, and $c = .2$ and the other with $a = .5$, $b = 5.0$, and $c = .2$. Despite the very large differences in *a* and *b* parameters, the two ICCs differ by less than .05 for abilities in the interval [−3, +3]. Thus, it is possible that an ICC computed from estimated item parameters could be very close to the ICC computed from actual item parameters despite large errors of estimation for *a* and *b*. If this is true, estimation accuracy should be studied by comparing recovered and actual ICCs. Positive results would suggest that IRT could be used for sample sizes and numbers of items much smaller than previously believed.

Method

Generation of Item Responses

Binary item response data were generated according to the two- and three-parameter logistic models (Birnbaum, 1968). The three-parameter logistic model represents the probability of a correct response to the i^{th} item as a function of three item parameters, item discrimination (a_i), item difficulty (b_i), and a pseudo-guessing parameter (c_i), and a single examinee parameter, ability (θ). Here

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} , \qquad [1]$$

where D is a scaling factor set equal to 1.7. The two-parameter model is the special case of the three-parameter model in which all c parameters equal zero.

The basic data for this study consisted of two 2,000 (examinees) by 60 (items) matrices of simulated binary item responses (1 = correct, 0 = incorrect). Both matrices were generated from common sets of abilities and item parameters. Ability values were sampled from a normal distribution with mean zero and unit standard deviation. The 60 a parameters were created by sampling numbers, X_i, from a uniform distribution in the interval [.3, 1.4] and then applying a 1.4 power transformation, $a_i = X_i^{1.4}$. The resulting a values had a mean of .862, a standard deviation of .209, and a positive skew, third moment = .233. The b values were sampled from a uniform distribution in the interval [−3, +3] and the c parameters were drawn from a uniform distribution in the interval [.11, .33].

Three-parameter logistic item responses were simulated by computing the probability of a correct response by Equation 1 for each of the 60 items and 2,000 values of θ. The probability of a correct response was then compared to a random number drawn from the [0, 1] uniform distribution. If the probability of a correct response was less than the sampled random number, the item was scored as incorrect; otherwise, the item was scored as correct. The Fortran random number generator in Spectrum IV (Taylor & Smith, 1976) was used. The two-parameter logistic item responses were created following the same procedure except that the c parameters were set equal to zero. All other item and person parameters remained the same.

Test Lengths and Sample Sizes

Sample sizes of N = 2,000, 1,000, 500, and 200 and test lengths of n = 60, 30, and 15 items were simulated. These sample sizes were chosen because they extended above the minimum size recommended by Lord for estimating item discrimination parameters and below the minimum that pretesting indicated would provide reasonable estimates of ICCs. The test lengths ranged from greater than the minimum suggested by Lord for parameter estimation to approximately those lengths studied by Swaminathan and Gifford (1979).

The 2,000 by 60 matrices were used to form four "tests" of length n = 15 items, two "tests" of length 30, and one of length 60. Items 1 to 15 formed the first test of n = 15 items, Items 16 to 30 formed the second, and so forth. Items 1 to 30 formed one 30-item test and the remaining 30 items were used as the second 30-item test.

The first 200, 500, and 1,000 simulated subjects from the total sample of 2,000 were used in analyses involving these sample sizes. This created a partial dependency in the results in that the sample of 200 was included in the sample of 500, and so forth. Supplementary analyses, however, indicated that any sample of 200 subjects drawn from the original 2,000 was approximately equivalent to any other sample of 200 in terms of the accuracy of recovered ICCs.

Item and person parameters were estimated by the CDC version of LOGIST. LOGIST's default convergence criteria were used throughout the present study. The maximum number of stages was set at 15 for each LOGIST run in order to conserve computer funds.

Evaluation of Estimation

Recovered ICCs were compared to actual ICCs calculated from the simulation parameters at 31 θ values chosen at equal intervals from −3.0 to +3.0 by the root mean squared error (RMSE) for each item (i):

$$\mathrm{RMSE} = \sqrt{\frac{1}{31}\sum_{j=1}^{31}[P_i(\theta_j) - \hat{P}_i(\theta_j)]^2}\ . \qquad [2]$$

The overall measure of ICC recovery was the average RMSE across all 60 items for a particular n and N combination. RMSEs were available for all 60 items even when n = 15, because four 15-item tests were analyzed by LOGIST.

Note that Equation 2 does not weight the root mean square by the expected frequency of individuals at various θ values. Thus, the calculated root mean squared errors do not refer to the differences between true and recovered ICCs evaluated at a set of θ values that would be expected in a typical sample of individuals. Instead, they refer to the distance between functions across a θ interval that encompasses most abilities encountered in practice. To the extent that items can be targeted to examinees so that the ability distribution has a mean near zero and a variance near unity, then the average RMSE provides a conservative index of recovery of ICCs.

Two measures of the accuracy of estimation of θ were computed. The first is the correlation or average correlation between θ and $\hat{\theta}$. In particular, θ and $\hat{\theta}$ were correlated for the 60-item tests. When $\hat{\theta}$'s were based on responses to 30- or 15-item tests, there were two and four estimates of θ, respectively. Here, the correlations were averaged (following an r to z transformation); thus, average $r_{\theta\hat{\theta}}$ values were obtained and not correlations between θ and averaged $\hat{\theta}$. RMSEs of $\hat{\theta}$ served as the second measure of estimation accuracy. RMSEs were computed for all $\hat{\theta}$'s available in each cell of the design.

Results

Recovery of ICCs

The results presented in Tables 1 and 2 and displayed graphically in Figures 1 and 2 indicate the recovery of two- and three-parameter logistic ICCs by LOGIST. The tables and figures present RMSEs averaged across 60 items.

Table 1 shows that the average RMSE for all combinations of 60- and 30-item tests with samples sizes of 500, 1,000, and 2,000 is less than .05 for the two-parameter logistic model. These errors indicate very accurate recovery of ICCs. With test lengths of 60 and 30 items, the average RMSE is less than .07 with as few as 200 subjects. For the 15-item test, samples of 2,000 and 1,000 resulted in average RMSEs slightly greater than .05. A sample size of 200 with only 15 items resulted in an average RMSE of nearly .09, which is large enough to cause serious concern.

The results are less impressive for the three-parameter logistic model (Table 2). For sample sizes of 2,000 and 1,000 and tests of 60 and 30 items, the average RMSEs are less than .05. For the longest test length evaluated, the errors are less than .06 with sample sizes of 500 and 200. The other combinations of sample sizes and test lengths resulted in RMSEs that were marginally larger to substantially larger.

LOGIST converged in 15 or fewer stages for 27 of 28 two-parameter logistic analyses. In contrast, convergence problems for the three-parameter logistic LOGIST analysis of 15-item tests were substantial. None of these LOGIST runs achieved convergence in 15 stages. Further

Table 1
Average Root Mean Squared Errors of Recovered Two-Parameter Logistic ICCs*

Number of Items	Sample Size			
	2000	1000	500	200
60	.022 (.012)	.028 (.018)	.041 (.025)	.068 (.036)
30	.030 (.017)	.036 (.025)	.045 (.034)	.069 (.037)
15	.052 (.022)	.054 (.027)	.065 (.037)	.088 (.043)

*Standard deviation of RMSEs in parentheses.

about 10 to 15 stages, with little apparent overall improvement. The second observation appears related to the first: average RMSEs of ICCs based on item parameter estimates obtained in 40 LOGIST stages were *not* generally smaller than the 15-stage average RMSEs shown in Figure 2. In fact, RMSEs were substantially larger for some 40-stage ICCs. Further remarks about these two observations are made below.

ICC Recovery Using a Decentered Ability Distribution

One possible reservation about these results is caused by the use in this study of a normal distribution of θ's with mean zero, which is the same as the mean of the b distribution. This particular distribution of θ was chosen because, in the absence of any strong a priori beliefs, a normal distribution of ability is a good approximation to the usually encountered distributions of ability measures. However, the centering of both distributions at zero may artificially reduce errors of estimation. To investigate the effects of a less optimally centered θ distribution, eight cells of the design (2,000 and 200 examinees crossed with 60 and 15 items, using the two- and three-parameter logistic models) were replicated using a normal (−.5, 1) distribution of ability. Following the procedures described above, subject and item parameters were estimated. Estimates of a and b were rescaled following the procedure described by Lord (1980b) so that they would be in a metric comparable to the true a and b values. RMSEs were computed by Equation 2 with θ values selected from −3.5 to 2.5 rather than from −3.0 to 3.0.[2] These results are shown in Table 3 for the two-parameter model and Table 4 for the three-parameter model.

The results for the decentered θ distribution are in general agreement with those based on the centered θ distribution. The largest difference occurs for three-parameter data, 2,000 subjects, and 15-item tests. The centered θ distribution yielded an average RMSE of only .068 in this cell, whereas the decentered distribution yielded an average RMSE of .160. Thus, in this cell of the design the location parameter of the ability distribution made a substantial difference in RMSE. This indicates a lower limit for estimating IRT parameters by LOGIST for badly decentered distributions of ability.

Estimation of Person Parameters

Table 5 presents the results of ability estimation in the 24 combinations of four sample sizes, three test lengths, and two test models for the centered ability distribution. The column headed "2PL" refers to item responses generated by the two-parameter logistic model and "3PL" refers to item responses generated by the three-parameter logistic model. Table 5 shows the number of finite ability estimates in each cell of the design.[3] Only finite values of $\hat{\theta}$ were used to compute the RMSEs and correlations between θ and $\hat{\theta}$ presented in Table 5.

There are a number of obvious trends in Table 5. For a fixed test length and item responses from the two-parameter logistic model, sample size has only a small influence on the accuracy of estimation of θ. Correlations between $\hat{\theta}_i$ and θ_i typically decreased by .02 from a sample size of 2,000 to a sample size of 200 and average RMSEs increased by .1. In contrast, the effect of decreasing test length on the accuracy of estimation of θ is pronounced. Halving the number of items from 60 to 30 resulted in a drop of .04 in the average $r_{\theta\hat{\theta}}$ of the two 30-item tests, and a further reduction to 15-item tests resulted in a decrease in the average correlations of approximately .10. Inspection of Table 5 also suggests that for two-parameter item responses, 30 items with any sample size down to 200 are sufficient for research in which θ is to be correlated with another variable.

[2] This interval was chosen because it was designed to compare true and recovered ICCs across an interval extending three standard deviations above and below the mean of the ability distribution. Other analyses indicated that average RMSEs computed for the interval (−3.0, +3.0) are very similar to the average RMSEs in Tables 3 and 4.

[3] The likelihood function is maximized at an infinite value of θ if all items are answered correctly or incorrectly.

Table 3
Average Root Mean Squared Errors of
Recovered Two-Parameter Logistic ICCs
for Decentered θ Distribution*

Number of Items	Sample Size	
	2000	200
60	.020	.048
	(.008)	(.015)
15	.061	.087
	(.025)	(.031)

*Standard deviation of RMSEs in parentheses.

The effects of sample size on the accuracy of estimates of θ are quite different for the three-parameter logistic model. Table 5 shows that decreases in $r_{\theta\hat{\theta}}$ and increases in average RMSE were substantial. There is a decrease in correlations as large as .18 and an increase in RMSE of more than 4.0 as sample size is reduced from 2,000 to 200. The effects of decreasing test length on accuracy of θ estimates are also profound. In fact, the very large RMSEs obtained for tests of length $n = 15$ show that maximum likelihood estimates of θ can be seriously biased: Simulated high ability examinees would occasionally "respond" correctly to all but one or two items and receive ability estimates as large as $\hat{\theta} = 88.11$.

Table 6 contains values of $r_{\theta\hat{\theta}}$ and RMSE computed for $\hat{\theta}$ values in the interval [−3, +3]. This interval contains the estimates of θ that would be credible to researchers; it is clear that $\hat{\theta}$ values of, say, 5, 15, or 30 should be interpreted only as evidence of high ability.[4] Restricting $\hat{\theta}$ to finite values in Table 5 eliminates up to 2% of the simulated examinees and the further restriction of $\hat{\theta}$ used to construct Table 6 eliminates up to an additional 7%. For 30- or 60-item tests, however, less than 3% of the simulated examinees are eliminated by this further restriction.

Values of $r_{\theta\hat{\theta}}$ show little change from Table 5 to Table 6 for the two-parameter logistic model. As expected, the RMSEs decrease, but the decreases are relatively small. In contrast, the re-

Table 4
Average Root Mean Squared Errors of
Recovered Three-Parameter Logistic ICCs
for Decentered θ Distribution*

Number of Items	Sample Size	
	2000	200
60	.042	.073
	(.018)	(.027)
15	.160	.174
	(.083)	(.079)

*Standard deviation of RMSEs in parentheses.

[4] It could be argued that the $\hat{\theta}$ interval used to construct Table 6 is still too wide. In this case, the correlations and average RMSEs in Table 6 are conservative estimates of correlations and average RMSEs that would be obtained if the $\hat{\theta}$ interval were further restricted.

Table 5
Root Mean Squared Errors of $\hat{\theta}$, Correlations Between θ and $\hat{\theta}$ and Number of Finite $\hat{\theta}$

	Number of Items					
Sample	60		30		15	
Size	2PL	3PL	2PL	3PL	2PL	3PL
N = 2000						
RMSE	.282	.487	.393	1.035	.560	2.450
$r_{\theta\hat{\theta}}$	.961	.905	.926	.727	.848	.578
Finite $\hat{\theta}$	2000	2000	4000	3996	7949	7898
N = 1000						
RMSE	.293	.504	.401	1.094	.557	3.194
$r_{\theta\hat{\theta}}$	.959	.898	.922	.695	.845	.552
Finite $\hat{\theta}$	1000	1000	2000	1997	3977	3950
N = 500						
RMSE	.313	.832	.411	1.428	.611	4.163
$r_{\theta\hat{\theta}}$	.956	.795	.918	.648	.834	.497
Finite $\hat{\theta}$	500	500	1000	998	1990	1976
N = 200						
RMSE	.382	1.254	.503	1.753	.662	6.505
$r_{\theta\hat{\theta}}$	.949	.724	.907	.593	.828	.430
Finite $\hat{\theta}$	200	200	400	398	792	785

sults for the three-parameter model are substantially changed in Table 6. The RMSEs are much smaller in this table than in Table 5. In Table 6 the three-parameter logistic average RMSEs are relatively constant for tests of a fixed length across all sample sizes: about .37 for n = 60, .55 for n = 30, and .80 for n = 15. The $r_{\theta\hat{\theta}}$ correlations are also relatively constant: about .93 for n = 60, .85 for n = 30, and .70 for n = 15.

Correlations Between Estimated and Actual Item Parameters

Although the emphasis of this study is on the recovery of ICCs, it is of interest to examine correlations between estimated and actual item parameters. These correlations are shown in Table 7.

The results summarized in Table 7 serve several purposes. First, there has been little previous monte carlo research on item parameter estimation for the two-parameter logistic model. Therefore, the correlations in Table 7 provide new information concerning accuracy of parameter estimation. Second, results for the three-parameter logistic model can be compared to previous work (e.g., Swaminathan & Gifford, 1979), which provides further insights into parameter estimation. Finally, conclusions drawn from Table 7 can be compared to conclusions drawn from Figures 1 and 2. For some purposes (e.g., determining minimum sample sizes and test lengths for practical applications of IRT) it appears that RMSEs of recovered ICCs provide the best index of item parameter estimation. If conclusions drawn from Table 7 differ from conclusions based on Figures 1 and 2, then practitioners may be misled if they examine only correlations between actual and estimated item parameters.

Columns 1 and 2 in Table 7 contain the correlations between true and estimated item pa-

Table 6
Root Mean Squared Errors of $\hat{\theta}$, Correlations Between θ and $\hat{\theta}$ and Frequency of $\hat{\theta}$ in -3 to +3 Interval

Sample Size	Number of Items 60 2PL	 3PL	 30 2PL	 3PL	 15 2PL	 3PL
N = 2000						
RMSE	.276	.377	.383	.529	.535	.744
$r_{\theta\hat{\theta}}$	.961	.928	.926	.860	.854	.739
Frequency	1970	1965	3981	3926	7921	7583
N = 1000						
RMSE	.278	.377	.397	.534	.545	.771
$r_{\theta\hat{\theta}}$	.961	.927	.920	.860	.845	.718
Frequency	994	987	1999	1967	3949	3785
N = 500						
RMSE	.289	.373	.398	.570	.577	.806
$r_{\theta\hat{\theta}}$	.958	.929	.915	.834	.834	.696
Frequency	495	496	994	970	1973	1881
N = 200						
RMSE	.290	.351	.423	.548	.613	.849
$r_{\theta\hat{\theta}}$	.962	.936	.907	.845	.821	.654
Frequency	198	197	396	390	782	731

rameters for the two-parameter logistic model. Although the correlations of b with $\hat{b}$ are all large and stable, the correlations are less lawfully behaved than the RMSEs. They do not display, for example, the effects of test lengths within a constant sample size. Correlations of a and $\hat{a}$ are smaller than correlations of b and $\hat{b}$, as is expected from past research. Again, these correlations are not as orderly as the average RMSEs shown in Figure 2. For example, $r_{a\hat{a}}$ is actually larger for the 15-item test than for the 60-item test (.908 vs. .920) for the $N = 2{,}000$ sample size. Since $r_{b\hat{b}}$ is virtually identical for the $n = 15$ and $n = 60$ item tests, it might be erroneously concluded that increasing test length *decreases* item parameter estimation accuracy. Figure 1 shows the substantial *increase* in estimation accuracy that actually occurs. Note that two items in the 60-item test with a sample of 200 did not converge and had final estimates of b that were less than −30 or greater than +30. These items were eliminated from consideration (as they would have been if they appeared in a test under development).

Columns 3 and 4 in Table 7 present the comparable results for the three-parameter logistic model. One three-parameter logistic item had a final $\hat{b}$ greater than 13 in all analyses and was therefore eliminated. Mirroring the results obtained for the RMSEs, these correlations are slightly to substantially lower than the corresponding correlations for the two-parameter model. Of greater interest, however, is the finding that the correlations in Table 7 are substantially lower than those obtained by Swaminathan and Gifford (1979). One difference between the studies is that b's were sampled from [−3, +3] in the present study, whereas Swaminathan and Gifford sampled b's from [−2, +2]. In columns 5 and 6 of Table 7 are the correlations between true and estimated a and b parameters for the 38 items in the present study whose true

Table 7
Correlations Between True and Estimated Item Parameters

Sample Size	Number of Items	2PL		3PL[b]		3PL[c]	
		$r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{a\hat{a}}$	$r_{b\hat{b}}$
2000	60	.908	.995	.567	.873	.835	.973
	30	.874	.981	.598	.901	.761	.988
	15	.920	.996	.623	.842	.685	.946
1000	60	.917	.992	.543	.939	.694	.992
	30	.897	.995	.537	.623	.706	.949
	15	.854	.994	.527	.746	.629	.861
500	60	.839	.983	.482	.812	.615	.866
	30	.827	.987	.446	.570	.578	.856
	15	.733	.984	.329	.585	.480	.728
200	60	.565[a]	.981[a]	.413	.664	.528	.970
	30	.623	.967	.446	.604	.514	.940
	15	.578	.941	.344	.765	.362	.721

[a] 58 items

[b] 59 items

[c] Correlations shown are based on the 38 items with $-2 < b_i < +2$.

item difficulties were between −2 and +2. Note, however, that these item parameters were estimated in tests of length n = 60, 30, and 15. The results in columns 5 and 6 indicate substantially greater estimation accuracy than suggested by columns 3 and 4 and also replicate very accurately earlier studies.

Discussion

The procedures adopted in this monte carlo study of parameter estimation in IRT were chosen to simulate an investigator working with scales and instruments that are not as highly developed as the SAT, ACT, GRE, WISC, or Stanford-Binet tests. Instead, the simulation design was constructed to be more similar to what the majority of psychological investigators can expect to encounter. Thus, items with b values more extreme than those recommended by Lord (1980b) and items with relatively low values of a were simulated. The decision was made to include a broad range of items in order to learn what IRT and LOGIST would do with items perhaps more typical of those encountered by practitioners and researchers.

The accuracy required of estimates of item and person parameters obviously depends on the questions being studied by the investigator. For example, in studies of item bias the emphasis is on accurate estimation of ICCs. Apparently, large numbers of items are not necessarily needed for these kinds of studies. Not surprisingly, however, large numbers of subjects are necessary. Test lengths as short as 30 items, if combined with sample sizes of 500 examinees for the two-parameter model or 1,000 for the three-parameter model, appear sufficient for accurate recovery of ICCs. Samples of 2,000 examinees, of course, yield more accurate estimates.

In many situations, emphasis is on accurate estimation of θ. The present research indicates that initial item calibration does *not* require large number of examinees for two-parameter logistic item responses. This conclusion is drawn because $r_{\theta\hat{\theta}} = .949$ and RMSE of $\hat{\theta}$ was .382 for a 60-item test with 200 simulated examinees. Thus, item parameters estimated from only $N = 200$ simulated examinees were accurate enough to provide very precise estimation of θ. Since large numbers of items are required for adaptive testing, the finding that calibration samples of 200 are sufficient for the two-parameter model should substantially decrease the cost of item pool development. A similar conclusion cannot be reached for the three-parameter logistic model. Estimating item parameters in samples of 500 and 200 simulated examinees reduced the accuracy of estimates of θ. This is shown in Table 5 by the increase in RMSE and decrease in $r_{\theta\hat{\theta}}$. In Table 6 the same effect is indicated by the decreased frequency of ability estimates in the range [−3, +3]. Apparently, Lord's recommendations concerning the number of subjects and items required for accurate ability estimation are well founded for the three-parameter logistic model.

The present results also suggest that there are tradeoffs between test length and sample size. Doubling test length and halving sample size, at least for tests of 30 and 60 items and sample sizes of 500, 1,000, and 2,000, resulted in comparable ICC average RMSEs. It is also clear that many different tests of different lengths with several replications per cell would be necessary to evaluate more carefully these tradeoffs. The costs of LOGIST analyses precluded evaluation of such tradeoffs.

The present research also illustrates the problems in estimating θ caused by nonzero lower asymptotes of ICCs. For fixed N and n, ability estimates tend to be more accurate with the two-parameter model than the three-parameter model. This is shown both by $r_{\theta\hat{\theta}}$ and RMSE of $\hat{\theta}$. In addition, there were many simulated examinees with $\hat{\theta}$'s that were excessively large in magnitude for the three-parameter model.

In three-parameter logistic analyses of 15-item tests, LOGIST failed to achieve convergence. In addition, there was a tendency for $\hat{\theta}$ values to become excessively large. Both of these difficulties indicate that the likelihood surface is not well suited to quadratic methods of function maximization. It seems reasonable to speculate that the difficulties are caused more by the likelihood surface than by the method of function maximization. If this is true, then the use of Bayesian estimation with informative priors may provide practically useful results. Further, use of informative prior distributions for person and item parameters appears to be justified because items are usually selected for a particular examinee population by a carefully designed process.

The results seen in Tables 5 and 6 indicate that long conventional tests are needed for very accurate estimation of θ. This underscores the potential benefits of adaptive testing. An effective item selection algorithm would choose items that are appropriately difficult or easy and consequently provide substantially more accurate estimation of ability on short tests.

No empirical study, monte carlo or otherwise, can provide compelling evidence about the consistency of an estimator of a parameter. The behavior of RMSEs for ICCs obtained from LOGIST is unknown if extrapolated beyond the limits of the sample sizes and test lengths presented in Figures 1 and 2. Nonetheless, the root mean squared errors, computed with equal weights along the θ continuum from −3 to +3 with only weak assumptions about the distribution of ability, suggest convergence to a trivially small error for both the two- and three-parameter logistic models. This is, of course, not conclusive evidence about the consistency of the estimators. Nonetheless, for most applications of IRT to real problems, these results suggest convergence.

References

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

Dorans, N., & Drasgow, F. Alternative weighting schemes for linear predictions. *Organizational Behavior and Human Performance,* 1978, *21,* 316-345.

Kendall, M., & Stuart, A. *The advanced theory of statistics* (Vol. 2, 4th ed.). New York: MacMillan, 1979.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. Item bias in a test of reading comprehension. *Applied Psychological Measurement,* 1981, *5,* 159-173.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement,* 1968, *28,* 989-1020.

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

Lord, F. M. Small N justifies Rasch methods. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980. (a)

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

Ree, M. J., & Jensen, H. E. Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Swaminathan, H., & Gifford, J. A. *Estimation of parameters in the three-parameter latent trait model* (Report No. 90). Amherst MA: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluation Research, 1979.

Taylor, F., & Smith, S. L. *Digital signal processing in Fortran.* Lexington MA: Lexington Books, 1976.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin,* 1976, *83,* 213-217.

Wood, R. L., & Lord, F. M. *User's guide to LOGIST* (Research Memorandum 76-4). Princeton NJ: Educational Testing Service.

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

Acknowledgments

We thank Professor Michael V. Levine for comments and suggestions made on earlier drafts of this manuscript. This research was supported, in part, by a University of Illinois Research Board computer time grant to Charles Hulin and by SSHRCC and FCAC doctoral fellowships to Robin Lissak. We thank Professor Emanuel Donchin for providing access to computing facilities in his laboratory.

Author's Address

Send requests for reprints or further information to Charles L. Hulin, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign IL 61820.