Application of Item Response Models to Criterion-Referenced Assessment

Ronald K. Hambleton
University of Massachusetts

Of interest in this study was the use of item response models for obtaining accurate examinee domain score estimates and for increasing the probabilities with which examinees are assigned correctly to mastery states with criterion-referenced test scores. Specifically, the purpose of this investigation was to compare the one-, two-, and three-parameter logistic test models for estimating domain scores and making mastery/nonmastery decisions. Computer simulation methods were used to recover a set of true domain scores with each of the logistic test models under a variety of testing conditions. Also, the percent of times the use of each model led to decisions which were consistent with decisions made with the true domain scores was studied. The one-parameter and three-parameter model resulted in highly comparable results for middle and high ability examinees, while for low ability examinees, the more general model always performed somewhat better.

The success of objectives-based programs in education, industry, and the military depends, to a considerable extent, upon the quality of the criterion-referenced tests that are used. Recent technical advances have made it possible to build high quality criterion-referenced tests (CRTs) as well as to carry out evaluations of the tests and the scores derived from them (see, for example, Berk, 1980; Hambleton, 1980; Popham, 1978). Still, several important problems remain. For example, methods are needed for obtaining more accurate examinee domain score estimates[1] and for increasing the probabilities with which examinees are assigned correctly to mastery states. Lengthening tests is a common method. However, it is often not practical to lengthen tests, particularly to the lengths they would need to be, to meet desirable levels for the reliability and validity of test scores (Wilcox, 1976).

Tailored/adaptive testing is a promising method, but at the present time it is somewhat impractical because it normally involves the administration of test items by a computer terminal (Lord, 1980b; Weiss, 1977). The use of Bayesian statistical procedures is another method for improving the accuracy of domain score estimates and mastery decisions (Novick & Jackson, 1974; Swaminathan, Hambleton, & Algina, 1975). This method does not require any changes in the usual way tests are administered. Improvements in measurement precision are attributable to the utilization of information ignored by non-Bayesian procedures: Bayesian procedures use not only the direct information provided by an examinee's test score but also collateral information contained in the item responses of other examinees and prior informa-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 1, Winter 1983, pp. 33–44

0146-6216/83/010033-12$1.60

[1]An examinee domain score is the proportion of items in a well-defined domain of content that he/she can answer correctly.

tion on other relevant data that are available on the examinees (e.g., test scores from other segments of a course). Bayesian methods can be criticized, however, because they use information which is not consistent with one of the underlying principles of criterion-referenced testing. The principle is that each examinee's performance should be assessed in relation to a set of competencies and independent of any prior beliefs about the examinee's performance and/or the test performance of other examinees.

Another possible solution is through the use of item response theory (IRT; Hambleton, 1979; Lord, 1980a; Wright & Stone, 1979). IRT appears promising for several reasons. First, the theory has been used successfully to address other psychometric problems such as test score equating and the study of item bias (Lord, 1980a; Linn, Levine, Hastings, & Wardrop, 1981). Second, the two- and three-parameter logistic test models incorporate additional information into the ability estimation process (by considering the discriminating power of the items and the likelihood of examinees answering items correctly by guessing). Third, while IRT incorporates several strong assumptions, the most important of the assumptions, unidimensionality, seems likely to be met with criterion-referenced tests because the content domains describing competencies of interest are not usually too heterogeneous or multi-faceted. Still, in spite of the variety of successful applications of IRT (see, for example, Hambleton, 1982), the success of particular applications of IRT to CRTs cannot be assured. It may be, for example, that the additional information provided by the two- and three-parameter logistic models is of modest value.

There are three common test score uses:

1. Scores obtained from the set of items in a test are used to rank order examinees,
2. Scores are used to make descriptive statements about examinee performance in relation to well-defined domains of content, and
3. Scores are used to make mastery/nonmastery decisions in relation to well-defined domains of content.

The first test score use is normally accomplished with norm-referenced tests; and for the other two uses, criterion-referenced tests are used. With respect to the first use, Hambleton and Cook (1980) carried out a comparative study of the one-, two-, and three-parameter logistic models under a variety of simulated testing conditions (for example, the statistical characteristics of item pools and test lengths were varied). They found that the three models provided highly comparable rankings for middle and high ability examinees. The three-parameter model led to substantially more valid rankings of examinees at the low end of the ability scale when guessing was allowed to influence test performance. However, there are no studies in the psychometric literature on the comparison of the three logistic models with respect to the two principal uses of criterion-referenced tests: to make descriptions and decisions.

The purpose of this study was to compare the one-, two-, and three-parameter logistic models for estimating domain scores (proportion-correct scores) and making mastery/nonmastery decisions. Attempts were made to recover a set of true domain scores with each of the three models under a variety of testing conditions. Also studied was the percent of times the use of each model led to decisions which were consistent with decisions made with the true domain scores. The study was conducted using computer simulation methods with tests of several lengths (10, 15, 20, and 40 test items) and several cutoff scores, and the results were compared at different levels of ability (ranging from very low to high). Examinee test performance was simulated with the three-parameter logistic model. Therefore, comparison of results from the one- and two-parameter models with the three-parameter model provided evidence concerning the robustness of the one- and two-parameter models.

Research Design

Domain Scores and Ability Scores

The relationship between domain scores and

ability scores is given by the test characteristic curve

$$\pi_i = \frac{1}{n} \sum_{g=1}^{n} P_g(\theta_i), \quad [1]$$

where

$P_g(\theta_i)$ is the probability associated with examinee i with ability level θ answering item g correctly,

n is the number of items in the test, and

π_i is the domain score (or relative true score) for examinee i on the items in the test.

The computer simulations were carried out on the θ metric, but for the purposes of evaluating results and comparing models, ability score estimates were transformed by the appropriate test characteristic curves to obtain corresponding domain score estimates. This transformation was carried out to maximize the interpretation of the results: The domain score scale is more familiar to criterion-referenced testers.

Variables Under Investigation

Test length. Test length was one of the key variables: The improvement in the precision of domain score estimates and decision accuracy for tests of increasing length was of interest. There was also interest in a comparison of test models at different test lengths. In two previous studies (Hambleton & Cook, 1980; Hambleton & Traub, 1973) it was found that with long tests there were relatively small differences among the models, whereas with short tests the differences were substantial. Tests of four lengths were considered: 10, 15, 20, and 40. Ability estimation is difficult with fewer than 10 items, and criterion-referenced tests with more than 40 items measuring a single objective are not common.

Sample size for item calibration. A large sample of simulated examinees ($N = 2{,}000$) was used to obtain item parameter estimates with each test model. The item pool consisted of 40 test items. The true item difficulty values (b values) were chosen in the range (−2.00, +2.00); item discrimination values (a values) were chosen in the range (.40, 2.00); and item pseudo-chance level values (c values) were chosen in the range (.15, .25). The choice of item parameters was made to be reflective of item parameters which have been obtained with real test data (Lord, 1968).

Selection of a cutoff score. Several cutoff scores were studied ($\theta_0 = -1.50,\ 0.00,\ 1.50$). These cutoff scores were then transformed to the corresponding domain score scale using the true parameters of the items in the test under study. Although this choice of cutoff score on the domain score scale may not be optimal in the sense of maximizing the probability of correct classifications with each test model, the cutoff method adopted in this study is common in many standard-setting studies.

Criteria for Evaluating the Results

The first criterion was the average absolute deviation associated with domain score estimates for a chosen examinee across N ($N = 200$) simulated test administrations,

$$\frac{\sum_{j=1}^{N} \left| \pi_i - \hat{\pi}_{ij} \right|}{N} \quad [2]$$

where π_i is given by Equation 1 and

$$\hat{\pi}_{ij} = \frac{1}{n} \sum_{g=1}^{n} \hat{P}_g(\hat{\theta}_{ij}). \quad [3]$$

In Equation 3, n is the number of items used in obtaining the ability estimate, $\hat{\theta}_{ij}$, for examinee i on the j^{th} replication of the test; and $\hat{P}_g$ is the probability associated with success on item g using the item parameter estimates associated with the estimation of $\hat{\theta}_{ij}$. The average absolute deviation between true and estimated domain scores was obtained with each test model for examinees at several ability levels. This criterion was appropriate for addressing the descriptive use of scores obtained with each test model.

A second criterion was a minor modification of the first. It involved working with deviations instead of absolute deviations in Equation 2. The second criterion permitted the study of bias associated with domain score estimates with each test model.

A third criterion was also used. It was possible to determine, once a cutoff score was specified, the percent of times an examinee's true and estimated domain scores (for the choice of items in the test) were in the same mastery category.

Computer Simulation Method

The seven steps below were followed in the computer simulations:

1. The characteristics of a "typical" pool of test items were specified:
   i. b uniformly distributed in the interval (−2.0, +2.0),
   ii. a uniformly distributed in the interval (.40, 2.0),
   iii. c uniformly distributed in the interval (.15, .25).
2. Item parameters (b, a, c,) were selected from the distributions above for 40 test items.
3. 2,000 examinees were drawn from a normal ability distribution with mean equal to zero and standard deviation equal to one.
4. The performance of 2,000 examinees was simulated on the 40-item test. (This step produces a 2000 × 40 matrix of item scores.) See Hambleton and Rovinelli (1973) for details on how examinee item responses were simulated.
5. The one-, two-, and three-parameter models were fitted to the data set so that item parameter estimates associated with each test model were obtained.
6. One value from each category below was selected:
   i. Examinee ability level (−2.00, −1.00, −.50, .50, 1.00),
   ii. test length (10, 15, 20, 40),
   iii. test model (1, 2, 3),
   iv. cutoff score (.50 above the chosen ability level or below),

   and, then, 200 response patterns for an examinee at the chosen ability level on the test were generated. When 10-item tests were needed, the first 10 items in the item pool were used; when 15-item tests were needed, the first 15 items were used; and so on. The response patterns were generated using the "true" item parameters from Step 2 and the ability level drawn from Step 6i. Using the estimated item parameters for the model under study, an ability estimate for each response pattern was obtained. Domain score estimates corresponding to the ability estimates were obtained with the test characteristic curve consisting of item parameter estimates for the model used to obtain the ability estimates. Finally, three statistics were calculated:

$$\frac{\sum_{j=1}^{200} \left| \pi_i - \hat{\pi}_{ij} \right|}{200}, \qquad [4]$$

$$\frac{\sum_{j=1}^{200} (\pi_i - \hat{\pi}_{ij})}{200}, \qquad [5]$$

and,

$$\begin{aligned} &P(\hat{\pi}_{ij} \geq \pi_0), \quad \text{if} \quad \pi_i \geq \pi_0, \\ &P(\hat{\pi}_{ij} < \pi_0), \quad \text{if} \quad \pi_i < \pi_0 . \end{aligned} \qquad [6]$$

   where π_o is the cutoff score on the π scale corresponding to the cutoff score, θ_0, on the θ scale.
7. Combinations of Steps 6i, ii, iii, and iv were used to simulate data, and then the results were used to compare the three logistic models.

Two general purpose computer programs were used. Examinee item response data were generated with a computer program prepared by Hambleton and Rovinelli (1973). Item and abil-

ity parameter estimation was carried out with LOGIST (Wood, Wingersky, & Lord, 1976).

Results

True and Estimated Item Parameters

The 40 true and estimated item parameters for the one-, two-, and three-parameter models are presented in Table 1. Table 2 provides a summary of the correlations among the true and estimated item parameters. LOGIST did an accurate job of recovering the true item parameters for the three-parameter model ($r_{b\hat{b}} = .983$, $r_{a\hat{a}} = .943$, $r_{c\hat{c}} = .511$). On the surface, it may appear that the c parameters were not properly estimated; but the relatively low correlation is at least in part due to the restricted range of true c parameters. The average absolute deviation between true and estimated c parameters was very small (about .023). On the other hand, it was distressing to observe the high number of test items with the same c-parameter estimate (28 of 40 items had the value .195). This finding was probably obtained because of the very small sample of simulated examinees at the lower end of the ability continuum (about 46 for $\theta < -2.0$).

The most revealing feature of the best two-parameter model estimates of the true three-parameter item characteristic curves is the high negative correlation between the item difficulty and discrimination estimates ($r_{\hat{b}\hat{a}} = -.643$). Such a relationship (positive or negative) in the item parameter estimates is highly undesirable because of the implications it has for test construction.

The impact of the shape of the ability distribution on the relationships among the true and estimated item parameters was studied by comparing the results obtained with a normal ability distribution and a uniform ability distribution. The relationships among the true and estimated item parameters obtained with the uniform distribution are also reported in Table 2. The results are essentially the same as results for the normal distribution, although the c parameters were somewhat better estimated when ability was uniformly distributed (the correlation increased from .511 to .667). The result is likely due to the increase in sample size at the lower end (below −2.00) of the ability distribution. There was an increase of about 287 examinees (from about 46 in the normal ability distribution to about 333 in the uniform ability distribution). The improvement in the estimation of c parameters was not achieved without a cost. With the shift in the ability distribution, there were fewer examinees left in the middle of the distribution to estimate the a parameters with as much precision. The correlation between the true and estimated item discrimination parameters dropped from .943 to .845.

Domain Score Estimation with Logistic Test Models

The results in Table 3 provide a basis for comparing the precision of domain score estimates at four test lengths with the one-, two-, and three-parameter models at five ability levels ranging from −2.0 to +1.0. Minor sampling errors aside, the table reveals that at low ability levels (−2.0 to −1.0) the three-parameter model provides substantially better domain score estimates than either the one- or two-parameter model. Apparently at the lower end of the ability continuum the one- and two-parameter models were not robust with respect to the deviations in the data set from the underlying assumptions concerning guessing behavior and variations in item discrimination. These two models were substantially more robust with longer tests. In fact, the three models produced highly comparable results for the longer tests considered in this investigation. For 10- and 15-item tests the improvements for low ability examinees were substantial. If the numbers reported in Table 3 are close to the true values, they suggest that for the item pool under consideration at a low ability level ($\theta = -2.0$) a test of approximately 40 test items with the one-parameter model would be needed to produce the same degree of precision as a 10-item test with the three-parameter

Table 1
Summary of True and Estimated Item Parameters
(N=2000, Normal Ability Distribution)

Item Number	Item Parameter Estimates								
	True			Three			Two		One
	b	a	c	b*	a*	c*	b**	a**	b***
1	-.762	.789	.222	-.870	.739	.195	-1.154	.661	-1.216
2	.051	1.515	.168	.069	1.832	.195	-.243	1.058	-.345
3	.817	1.296	.180	.816	1.557	.214	.673	.511	.414
4	1.754	1.602	.152	1.767	1.988	.159	3.777	.240	1.502
5	.714	.681	.177	.813	.636	.195	.549	.322	.190
6	-.415	1.384	.224	-.504	1.464	.195	-.749	1.252	-1.062
7	-.557	.522	.195	-.568	.524	.195	-1.086	.397	-.824
8	.820	.968	.226	.819	1.092	.229	.587	.402	.270
9	.669	1.469	.189	.625	1.760	.195	.439	.669	.314
10	1.733	1.866	.225	1.636	1.988	.202	3.010	.238	1.172
11	.740	1.843	.223	.702	1.988	.218	.522	.584	.342
12	.094	1.974	.162	.037	1.988	.140	-.170	1.388	-.237
13	1.346	.534	.205	1.311	.542	.195	1.276	.229	.426
14	-.693	1.017	.226	-.723	1.170	.195	-.951	1.063	-1.282
15	-1.935	1.356	.170	-2.271	1.126	.195	-1.608	1.991	-2.986
16	1.590	1.121	.192	1.583	1.753	.220	2.573	.245	1.015
17	-1.252	.533	.176	-1.007	.648	.195	-1.316	.581	-1.268
18	1.670	1.251	.207	1.746	.883	.195	2.734	.220	.985
19	-1.908	1.520	.186	-2.068	1.320	.195	-1.580	1.991	-2.921
20	.366	1.062	.212	.411	1.253	.230	.023	.599	-.084
21	1.542	1.620	.168	1.450	1.650	.153	2.512	.323	1.258
22	.250	.904	.165	.311	1.092	.195	-.007	.610	-.109
23	.291	1.499	.227	.222	1.699	.195	-.094	.852	-.176
24	-.009	1.340	.184	-.000	1.480	.195	-.331	.926	-.447
25	1.309	1.232	.164	1.303	1.247	.144	1.958	.368	1.071
26	1.155	1.797	.226	1.147	1.988	.225	1.357	.378	.729
27	.590	1.874	.195	-.558	1.988	.190	.358	.730	.263
28	-1.755	1.786	.183	-1.909	1.988	.195	-1.660	1.991	-3.107
29	-1.269	1.255	.224	-1.330	1.425	.195	-1.201	1.993	-2.066
30	.249	.427	.178	.294	.432	.195	-.348	.254	-.257
31	.053	.771	.180	-.019	.716	.195	-.439	.489	-.450
32	-1.536	1.487	.235	-1.608	1.449	.195	-1.368	1.991	-2.446
33	.058	1.031	.224	.020	1.024	.195	-.337	.674	-.420
34	-.255	.532	.208	-.311	.520	.195	-.883	.359	-.643
35	-.580	1.803	.202	-.555	1.988	.195	-.771	1.742	-1.196
36	-.905	.569	.223	-.963	.597	.195	-1.323	.516	-1.177
37	.244	1.236	.180	.288	1.422	.195	-.010	.727	-.100
38	-1.307	1.052	.214	-1.438	.923	.195	-1.366	1.122	-1.919
39	-.034	1.069	.228	-.119	1.082	.195	-.463	.751	-.565
40	.033	.867	.244	-.071	1.003	.195	-.425	.678	-.507

Table 2
Intercorrelations Among Item Parameters and Item Parameter Estimates from the One-, Two-, and Three-Parameter Logistic Test Models (N=2000)[a]

Variable	True			Item Parameter Estimates: Three			Two		One
	b	a	c	b*	a*	c*	b**	a**	b***
b		.132	-.138	.983	.225	-.055	.929	-.787	.981
a	.132		-.075	.064	.943	-.171	.266	.396	.079
c	-.138	-.075		-.145	.098	.511	-.218	.059	-.166
b*	.994	.098	-.158		.170	-.039	.913	-.797	.969
a*	.049	.845	-.011	.033		-.079	.331	.288	.187
c*	-.038	-.211	.667	-.011	-.119		-.159	-.076	-.093
b**	.973	.154	-.248	.981	.091	-.100		-.643	.887
a**	-.729	.354	.034	-.743	.466	-.230	-.728		-.837
b***	.989	.117	-.251	.987	.041	-.138	.979	-.714	

[a]In the upper triangle are the intercorrelations when a normal distribution of ability was used. In the lower triangle are the correlations when a uniform distribution of ability was used.

model. Also, for moderate and high ability levels (−.50 to 1.0) the three models provide highly comparable domain score estimates.

Bias in Domain Score Estimation

Table 4 provides a summary of the bias in domain score estimation for four test lengths and nine ability levels with the three logistic models. On average, the bias in domain score estimation is very small. There is a definite trend in the results, however. At the low end of the scale, domain score estimates are, on the average, slightly overestimated. The reverse is true at the higher end of the scale. Thus, even though the data were generated to be consistent with the assumptions of the three-parameter model, ability estimates (and associated domain score estimates) obtained with the best fitting item parameter estimates for the one- and two-parameter models were only slightly biased.

Decision Accuracy with Logistic Test Models

Table 4 also contains information showing decision accuracy for the three logistic test models at four test lengths and five ability levels. Several points are of interest. First, the two-parameter model decision accuracy results are somewhat unstable. Probably the result was to be expected, given the poor item parameter estimates reported in Tables 1 and 2. Second, while there are a few minor reversals (probably due to sampling errors), the one-parameter model functioned about as well as the three-parameter model, except at the low end of the ability continuum

Table 3
Domain Score Estimation and Decision Accuracy for Several Ability Levels, Standards, and Test Lengths with Three Logistic Test Models

Ability Level	Standard (θ_0)	Number of Test Items	Test Model	Average Absolute Deviation[a]	Decision Accuracy[b]
-2.00	-1.50	10	1	.114	.61
			2	.101	.64
			3	.057	.63
	-1.50	15	1	.091	.64
			2	.085	.72
			3	.061	.68
	-1.50	20	1	.083	.65
			2	.072	.71
			3	.053	.78
	-1.50	40	1	.054	.84
			2	.047	.86
			3	.038	.88
-1.00	-1.50	10	1	.102	.61
			2	.100	.54
			3	.071	.67
	-1.50	15	1	.091	.65
			2	.071	.70
			3	.060	.79
	-1.50	20	1	.078	.72
			2	.060	.80
			3	.055	.80
	-1.50	40	1	.055	.82
			2	.044	.91
			3	.039	.93
-0.50	0.00	10	1	.116	.76
			2	.108	.81
			3	.091	.81
	0.00	15	1	.094	.77
			2	.075	.83
			3	.068	.84
	0.00	20	1	.079	.83
			2	.065	.83
			3	.063	.84
	0.00	40	1	.054	.90
			2	.046	.94
			3	.045	.92

-continued-

Table 3 (continued)

Ability Level	Standard (θ_0)	Number of Test Items	Test Model	Average Absolute Deviation	Decision Accuracy
0.50	0.00	10	1	.112	.80
			2	.107	.81
			3	.105	.80
	0.00	15	1	.097	.89
			2	.082	.85
			3	.082	.83
	0.00	20	1	.078	.89
			2	.073	.93
			3	.073	.92
	0.00	40	1	.050	.96
			2	.050	.96
			3	.048	.97
1.00	1.50	10	1	.101	.77
			2	.091	.77
			3	.095	.81
	1.50	15	1	.095	.79
			2	.072	.81
			3	.076	.80
	1.50	20	1	.068	.81
			2	.061	.86
			3	.060	.84
	1.50	40	1	.046	.94
			2	.043	.96
			3	.044	.96

[a]Average absolute deviation is obtained by comparing an examinee's domain score (corresponding to his/her ability level) to 200 independent estimates of his/her domain score.

[b]Decision accuracy is determined by calculating the proportion of times an examinee is classified on the basis of domain score estimates into the correct mastery state (determined by comparing the examinee's domain score to the chosen standard).

(−2.0 to −.5). Although not reported here (and not surprising), the differences between the one- and three-parameter models were even greater for low ability examinees when they were closer than .50 to the cutoff score. For average and high ability estimates, the one- and three-parameter models essentially led to the same rates of decision accuracy.

Conclusions

Several reasons have been offered in the measurement literature for not using the two- and three-parameter logistic models:

1. They require too much computer time for parameter estimation.

Table 4
Analysis of Bias[a] in Domain Score Estimation
(N=200)

Test Length	Model	θ -2.00	-1.50	-1.00	-0.50	0.00	0.50	1.00	1.50	2.00
10	1	-.010	-.011	-.015	-.015	-.015	-.005	-.002	.005	.006
	2	.004	.007	.002	-.005	-.019	-.008	.001	.024	.052
	3	-.025	-.018	-.007	.001	-.010	.007	.011	.025	.026
15	1	-.005	-.001	.002	-.003	-.003	-.012	-.011	-.016	-.017
	2	-.003	-.007	-.002	-.003	-.003	-.006	.001	.019	.046
	3	-.023	-.017	-.005	.002	.006	.010	.016	.046	.008
20	1	.001	-.002	-.004	-.009	-.008	-.009	-.003	-.004	-.001
	2	-.005	-.011	-.009	-.007	.001	.003	.019	.032	.052
	3	-.006	-.002	.003	-.001	.006	.009	.022	.022	.020
40	1	-.007	-.008	-.009	-.007	-.009	-.006	-.005	-.002	-.001
	2	.007	-.003	-.008	-.008	-.008	-.006	-.005	.004	.024
	3	.005	.001	-.003	.004	.006	.005	-.005	.001	.003

[a]Average difference, $\pi - \hat{\pi}$. Each entry in the table is based upon 200 estimates of the difference.

2. The computer program in common use (LOGIST) places strict constraints on the *a* and *c* parameters to obtain convergence.
3. The *c* parameters are poorly estimated.

A study by Hutten (1981) shows clearly that the computer costs are not unreasonable. Across 25 data sets with an average of 1,000 examinees and 50 test items, her average cost per data set for CPU time (at a rate of $800/hour) was $69.00.[2] While parameter constraints are used in the parameter estimation algorithm, a not uncommon technique in parameter estimation, the derived three-parameter model estimates from the computer program (LOGIST) were quite acceptable in this study (see, also, Lord, 1975). Also, from the results reported here and from Lord (1975), it can be seen that when there are a substantial number of low-ability examinees in the sample, the *c* parameters can be properly estimated. In addition to the above results, the principal results in this study suggest that when making domain score estimates and/or mastery/nonmastery decisions, the three logistic models give similar results except at the low end of the ability continuum where the three-parameter logistic model does a substantially better job. The more general models do function a little better overall, but probably not enough to justify their use in most classroom settings for assessing examinee ability. On the other hand, when sample sizes are large enough to obtain accurate item parameter estimates for the three-parameter model, and when there is special interest in examinees at the lower end of the ability continuum (for example, as there might be when identifying examinees in need of remediation or when assessing examinee abilities on a pretest administration), there appear to be substantial advantages for using the three-parameter model. It is, of course, important to stress that this conclusion should not be generalized to other applications of item response models.

Finally, it should be stressed that the results in this paper do not address the utility of item response models for use with actual criterion-referenced test data. It remains to be seen how well any of the item response models fit criterion-referenced test data. In this simulation study a comparison of the one-, two-, and three-parameter logistic test models was carried out with data sets which were consistent with the assumptions of the three-parameter logistic model. For this reason the study may be viewed as a study of the robustness of the one- and two-parameter models.

[2]These figures will seem low to users of LOGIST on IBM equipment. It should be noted that IBM uses a 32-bit word; and in order to get floating point precision, two of those words are used to make a 64-bit word, whereas CDC equipment has a 60-bit word. Therefore, IBM is slower because more code is executed to obtain similar precision with CDC. Also, the CDC code processing unit is faster than IBM equipment. Finally, CDC has a special hardware feature which allows it to execute instructions handling integer numbers and floating-point numbers at the same time. On IBM equipment each instruction is handled separately. The author is grateful to Richard Rovinelli for researching several strengths and weaknesses of IBM and CDC equipment.

References

Berk, R. (Ed.). *Criterion-referenced measurement: The state of the art.* Baltimore MD: The Johns Hopkins Press, 1980.

Hambleton, R. K. Latent trait models and their applications. In R. Traub (Ed.), *Methodological developments: New directions for testing and measurement* (No. 4). San Francisco: Jossey-Bass, 1979.

Hambleton, R. K. (Ed.), Contributions to criterion-referenced testing technology. *Applied Psychological Measurement,* 1980, *4,* 421-581.

Hambleton, R. K. (Ed.), *Applications of item response theory.* Vancouver BC: Educational Research Institute of British Columbia, 1982.

Hambleton, R. K., & Cook, L. L. The robustness of latent trait models and effects of test length and sample size on the precision of ability estimates. 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.

Hambleton, R. K., & Rovinelli, R. A FORTRAN IV program for generating examinee response data from logistic test models. *Behavioral Science,* 1973, *17,* 73-74.

Hambleton, R. K., & Traub, R. E. Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 195–211.

Hutten, L. *Fitting the one- and three-parameter models to empirical data.* Unpublished doctoral dissertation, University of Massachusetts, Amherst, 1981.

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* (RB-75-33). Princeton NJ: Educational Testing Service, 1975.

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

Lord, F. M. Some how and which for practical tailored testing. In L. J. Th. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates.* New York: Wiley, 1980. (b)

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1974.

Popham, W. J. *Criterion-referenced measurement.* Englewood Cliffs NJ: Prentice-Hall, 1978.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 1975, *12*, 87–98.

Weiss, D. J. (Ed.). *Applications of computerized adaptive testing* (Research Report 77-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Wilcox, R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1976, *1*, 359–364.

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.

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

Acknowledgments

The author is grateful to Craig Mills for preparing the necessary computer programs and for carrying out the computer analyses. The paper was finished while the author was on sabbatical leave at the University of Leyden, The Netherlands. A complete report of this study appears in Laboratory of Psychometric and Evaluative Research Report No. 117. *Amherst, MA: School of Education, University of Massachusetts, 1981. This research was performed pursuant to Contract F33615-79-C-0020 from the U.S. Air Force Human Resources Laboratory. However, the opinions do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred.*

Author's Address

Send requests for reprints or further information to Ronald K. Hambleton, University of Massachusetts, Laboratory of Psychometric and Evaluative Research, Hills South, Room 152, Amherst MA 01003.