DOCUMENT RESUME

ED 161 894 TM 007 324

AUTHOR Hambleton, Ronald K.; Cook, Linda L.
TITLE Some Results on the Robustness of Latent Trait Models.
PUB DATE [Mar 78]
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Complexity Level; Goodness of Fit; *Item Analysis; Scoring; Simulation; *Test Construction; Testing; Test Items
IDENTIFIERS Birnbaum 3 Parameter Logistic Model; *Item Characteristic Curve Theory; Item Discrimination (Tests); *Latent Trait Theory; Test Length; Test Theory

ABSTRACT

The purpose of the present research was to study, systematically, the "goodness-of-fit" of the one-, two-, and three-parameter logistic models. We studied, using computer-simulated test data, the effects of four variables: variation in item discrimination parameters, the average value of the pseudo-chance level parameters, test length, and the shape of the ability distribution. Artificial or simulated data representing departures of varying degrees from the assumptions of the three-parameter logistic test model were generated and the "goodness-of-fit" of the three test models to the data was studied. From the data sets analyzed in the study, it is clear that there are some sizable gains to be expected with modest length tests (n=20) in the correct ordering of examinees at the lower end of the ability continuum when three-parameter model estimates are used (as opposed to the number right score). The gains were cut roughly in half when the tests were doubled (n=40) in length. Item discrimination parameters as scoring weights had very little effect on the results. (Author)

Some Results on the Robustness of Latent Trait Models[1,2,3]

Ronald K. Hambleton and Linda L. Cook
University of Massachusetts, Amherst

-Printed in the U.S.A.-

[1] Paper presented at the annual meeting of AERA, Toronto, 1978.

[2] Laboratory of Psychometric and Evaluative Research Report No. 69. Amherst, MA: School of Education, University of Massachusetts, Amherst, 1978.

[3] This project was performed pursuant to a contract from the United States Office of Scientific Research. However, the opinions expressed here do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred.

Some Results on the Robustness of Latent Trait Models

Ronald K. Hambleton and Linda L. Cook
University of Massachusetts, Amherst

Abstract

The purpose of the present research was to study, systematically, the "goodness-of-fit" of the one-, two-, and three-parameter logistic models. We studied, using computer-simulated test data, the effects of four variables: Variation in item discrimination parameters, the average value of the pseudo-chance level parameters, test length, and the shape of the ability distribution. Artificial or simulated data representing departures of varying degrees from the assumptions of the three-parameter logistic test model were generated and the "goodness-of-fit" of the three test models to the data was studied.

From the data sets analyzed in the study, it is clear that there are some sizable gains to be expected with modest length tests (n = 20) in the correct ordering of examinees at the lower end of the ability continuum when three-parameter model estimates are used (as opposed to the number right score). The gains were cut roughly in half when the tests were doubled (n = 40) in length. Item discrimination parameters as scoring weights had very little effect on the results.

The topic of latent trait theory was introduced to educational measurement specialists over 25 years ago by Frederic Lord (1952, 1953). Until recently, his work and the work of other psychometricians in the latent trait theory field received only limited attention from test practitioners. However, important breakthroughs recently in problem areas such as test score equating, tailored testing, test design and test evaluation through applications of latent trait theory have attracted considerable interest from measurement specialists. Other factors that have contributed to the current interest in latent trait theory include the availability of a number of useful computer programs, publication of a variety of successful applications in measurement journals, and the strong endorsement of the field by authors of the last three reviews of test theory in the Annual Review of Psychology. Another testimony to the current interest and popularity of the topic is the fact that the Journal of Educational Measurement published six invited papers on latent trait theory and applications in the summer issue of 1977. (See for example, Hambleton and Cook, 1977; Lord, 1977; Wright, 1977.)

All of the latent trait models of interest in this paper (the one-, two-, and three-parameter logistic test models) rest on one important assumption: For practical reasons it is usually assumed the items are homogeneous in the sense they measure the same single ability. From there, users must specify the mathematical form of the "item characteristic curves." An item characteristic curve represents the probability of a correct answer to an item expressed as a function of ability. In the one-parameter model, items may vary in difficulty levels only; in the two parameter model, items may vary both in level of difficulty and discrimination; and in the three-parameter model, itemsmay vary in level of difficulty, discrimination, and pseudo-chance levels. The mathematical form of the three-parameter logistic curve is written

$$P_g(\theta) = c_g + (1-c_g)\,\frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}}\ ,\ g = 1,\ 2,\ \ldots,\ n.$$

In this expression, $P_g(\theta)$ is the probability that an examinee with ability θ answers item g correctly; "b_g" is the index of item difficulty; "a_g" is the index of item discrimination; and "c_g" is the pseudo-chance level. The reader is referred to Hambleton and Cook (1977) for a more detailed discussion of the item parameters. It should be noted that the item characteristic curves can be applied to binary scored items administered under non-speeded test conditions. The two-parameter model is obtained from the three-parameter model by setting c_g=0. The one-parameter model is obtained from the three-parameter model by setting c_g = 0 and a_g = a constant, g = 1, 2, ..., n.

While the potential usefulness of latent trait models is high, there remain many practical problems to address at the application stage. For one, how does a user go about selecting a latent trait model? One might be tempted to say that the user should always work with the more general models since these models will provide the "best" fits to the available test data. Unfortunately, the more general latent trait models (for example, the three-parameter logistic test model) require more computer time to obtain satisfactory solutions, require larger samples of examinees and longer tests, and are more difficult for practitioners to work with. Clearly, more needs to be known about the "goodness-of-fit" and "robustness" of latent trait models. Such information would aid practitioners in the important step of selecting a test model.

There has been some work on the "goodness-of-fit" between latent trait models and a variety of test data sets (see for example, Lord, 1975; Tinsley and Dawis, 1977; and Wright, 1968) and generally the results have been good (Hambleton, Swaminathan, Cook, Eignor, and Gifford, 1978). Only one study we have seen compared the fit of more than one latent trait model to the same test data sets (Hambleton and Traub, 1973). In this study, improvements were obtained in predicting test score distributions (for three tests) from the two-parameter model as compared to the one-parameter model.

On the question of model robustness (i.e., the extent to which the assumptions underlying the test model can be violated to a greater or lesser extent by the test data and be "fitted" by the model), the results of several studies have been reported (Dinero and Haertel, 1977; Hambleton, 1969; Hambleton and Traub, 1976; Panchapakesan, 1969). The results have been mixed, perhaps because of the confounding of results with sample sizes.

The problem as we see it with most of the goodness-of-fit studies and the robustness studies reported to date is that they provide no indication of the practical consequences of fitting a "less than perfect" model to a test data set. It really is of little interest to the practitioner to know that 15 out of 20 items failed to be fitted by a test model when the range of discrimination parameters reached (say) a value of .80. For one thing, if the size of the examinee sample is large enough, probably all items could be identified by a chi-square statistic of goodness-of-fit as not fitting the model. If the size of the examinee sample is small enough, perhaps none of the items would be misfit by the model! We think it would be interesting for practitioners to see comparisons of latent trait models and then "fit" to various data sets using a criterion measure (or measures) that have some practical meaning to them. To date there have been no comparative studies of the various latent trait models using practical criteria to judge the results.

Purposes of the Research

The purpose of the present research was to study, systematically, the "goodness-of-fit" of the one-, two-, and three-parameter logistic models. We studied, using computer-simulated test data, the effects of four variables: Variation in item discrimination parameters, the average value of the pseudo-chance level parameters, test length, and the shape of the ability distribution. Artificial or simulated data representing departures of varying degrees from the assumptions of the three-parameter logistic test model were generated and the "goodness-of-fit" of the three test models to the data was studied.

How should "goodness-of-fit" be measured? It seemed to us that, in some testing situations, (for example, some situations involving norm-referenced tests), test users desire to rank examinees based on their test score performance in a way that will closely reflect rankings based on examinee "true ability." Much effort is made by test developers to rank examinees properly (i.e., "validly") by using suitably long tests, high-quality test items, proper test conditions and so on. Utilizing the two- and three-parameter models with many test data sets will also be helpful in accomplishing the stated goal of ranking examinees in a way that will be consistent with rankings based on "true" ability scores.

In this study, because we used simulated data, it was possible to "know" examinee ability scores. They served as our criterion against which to judge the statistics derived from the three test models for ranking examinees. Three statistics, derived from the one-, two-, and three-parameter logistic models, respectively, were obtained and used to rank examinees. The rankings of examinees derived from each model (for each set of test data) were then compared to examinee "true" abilities. The Spearman rank difference formula was used to summarize the similarity between each pair of ranks (true abilities and estimates of ability from one of the models). We also reported the average size of the discrepancies in the ranks for each group of 500 examinees.

As an aside, we note that it would have been desirable also to compare ability estimates, denoted $\hat{\theta}$, and true ability scores, denoted θ. Unfortunately, because of the arbitrariness of the scale on which $\hat{\theta}$ is measured, it would have been of very limited value to report summary statistics such as $\sum_{i=1}^{N} |\theta_1 - \hat{\theta}_1|/N$. In some of our later work we will address the scaling problem through equating methods.

Method

Simulating the Test Data

The simulation of item response data for examinees was accomplished using the three-parameter logistic model. First, the number of examinees (N), shape of the ability distribution, and values of the ability parameters (θ_i = 1, 2, ..., N) were specified. Next, the number of items in the test (n) and values of the three item parameters (a_g, b_g, c_g, g = 1, 2, ..., n) were specified. Then the examinee and item parameters were substituted in the equation of the three-parameter logistic model to obtain a number p_{ij} $(0 \leqslant p_{ij} \leqslant 1)$ representing the probability that examinee i correctly answered item j. The probabilities were arranged in a matrix P of order Nxn whose (i, j)th element was p_{ij}. P was then converted into a matrix of the item scores for examinees (1 = correct answer, 0 = incorrect answer) by comparing each p_{ij} with a random number obtained from a uniform distribution on the interval [0, 1]. If the random number was less than or equal to p_{ij} (which would happen on the average p_{ij} of the time), p_{ij} was set equal to 1, otherwise p_{ij} was set to 0. The matrix P of zeros and ones was the simulated test data. At this point, three statistics used in estimating examinee ability were calculated: $\sum_{g=1}^{n} u_g$, $\sum_{g=1}^{n} a_g u_g$, and $\sum_{g=1}^{n} w_g(\theta)\, u_g$, corresponding to statistics which are used in the estimation of examinee ability with the one-, two-, and three-parameter models, respectively. (Recall, $u_g = 1$ for a correct response, $u_g = 0$, otherwise.) For the three-parameter model statistic, since the item weights $[w_g(\theta)]$ depend on examinee ability, we obtained three-parameter model estimates of ability for each examinee from LOGIST (Wood, Wingersky, and Lord, 1976).[1] Once we had calculated the three-parameter model estimates of ability, we use them (instead of $\sum_{g=1}^{n} w_g(\theta)\, u_g$) for convenience.

[1] There has been some discussion by practitioners of the difficulties of using LOGIST, and the costs involved. We were able to install the program very quickly on our CYBER 70 System and the cost of typical runs in our study (20 or 40 items, 500 examinees) was about $2.00. We should add that these results were obtained for the case where item parameters are known.

The values of the examinee and item parameters were chosen as follows:

Examinee Parameters. The number of examinees was set equal to 500. This number was sufficient to produce stable goodness-of-fit results. Two distributions of ability were considered: Uniform [-2.5, 2.5] and Normal [0, 1].

Item Parameters. Two test lengths (20 and 40 items) were used in the simulations. Both values are fairly typical of test lengths in common use.

In the simulation of test data, item difficulty parameters, b_g, $g = 1, 2, \ldots, n$, were selected at random from a uniform distribution on the interval [-2, 2]. An analysis of the difficulty parameters reported by Lord (1968) suggested that this decision was reasonable.

The discrimination parameters, a_g, $g = 1, 2, \ldots, n$, for the items of a simulated test were selected at random from a uniform distribution with mean = 1.12. The range of the discrimination parameters was a variable under investigation. The range was varied from 0.0 to a maximum of 1.24 [.50 to 1.74], and an intermediate value of .62 [.81 to 1.43] was also studied. The maximum value of discrimination was similar to the range and distribution of the discrimination parameters reported for the Verbal Section of the SAT (Lord, 1968).

The extent of guessing in the simulated test data was another variable under study. Two values of the average guessing parameter were considered: $\bar{c} = 0.00$, and $\bar{c} = 0.25$. All pseudo-chance level parameters were set equal to the mean value of the c-parameter under investigation.

Factor Structure. For all of the tests simulated in the study, it was assumed that the test items were unidimensional, i.e., measured a common trait.

Goodness-of-Fit

The approach to goodness of fit was described earlier in the purposes section of the paper. For each data set (24 in total; 2 test lengths x 2 levels of pseudo-chance parameters x 3 levels of variation in discrimination parameters x 2 ability distributions), three statistics used in estimating ability for the one-, two-, and three-parameter models, respectively, were calculated and compared to the true ability parameters. Comparisons were made via the use of Spearman rank difference formula and the average discrepancy in ranks.

To further facilitate the interpretation of results, they are reported separately for each half of the ability distribution as well as for the total ability distribution.

Results

The results of our computer simulations are summarized in Tables 1 to 6. The first row of each table was inserted to serve as a check on our calculations.

For convenience we will discuss the results in point form around the variables under study:

Level of Variation in Discrimination Parameters

1. For the values studied in the paper, using discrimination parameters as item weights contributed very little to the proper ranking of examinees.

Level of Pseudo-Chance Level Parameters

2. With the twenty-item tests; the three-parameter model was considerably more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .08 higher (∿.75 to ∿ .83) in the uniform distribution of ability and about .08 higher in the normal

Table 1

Summary of the Goodness-of-Fit Results
(Uniform Ability Distribution,[1] $\theta = -2.5$ to 0.0)

					Comparison of Estimates					
					True Versus One Parameter Model		True Versus Two Parameter Model		True Versus Three Parameter Model	
Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	SD	r[2]	AAD[3]	r	AAD	r	AAD
20	0.00	.00	5.03	3.00	.881	54.238	.881	54.238	.881	54.238
20	0.00	.25	8.98	2.86	.765	76.610	.765	76.610	.827	64.984
20	.81 to 1.43	.00	5.24	3.10	.877	56.068	.876	56.406	.876	56.404
20	.81 to 1.43	.25	9.01	2.84	.760	77.144	.764	76.900	.833	64.284
20	.50 to 1.74	.00	5.36	3.02	.874	56.496	.874	56.558	.874	56.562
20	.50 to 1.74	.25	9.12	2.83	.747	80.076	.750	79.920	.827	65.770
40	0.00	.00	9.58	6.22	.944	36.482	.944	36.482	.944	36.482
40	0.00	.25	17.82	5.33	.868	58.578	.868	58.578	.908	48.704
40	.81 to 1.43	.00	10.14	6.37	.949	36.504	.949	36.474	.949	36.474
40	.81 to 1.43	.25	17.98	5.41	.872	57.662	.875	56.860	.912	48.014
40	.50 to 1.74	.00	9.97	6.39	.942	37.862	.946	36.962	.946	36.742
40	.50 to 1.74	.25	18.18	5.41	.870	57.824	.876	56.872	.910	48.222

[1] N = 500

[2] Spearman Rank-Difference Formula

[3] Average absolute difference in rank order

Table 2

Summary of the Goodness-of-Fit Results
(Uniform Ability Distribution,[1] θ = 0.00 to +2.5)

					Comparison of Estimates					
					True Versus One Parameter Model		True Versus Two Parameter Model		True Versus Three Parameter Model	
Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	SD	r^2	AAD[3]	r	AAD	r	AAD
20	0.00	.00	14.99	2.82	.883	54.450	.877	55.624	.877	55.624
20	0.00	.25	16.21	2.13	.835	63.676	.828	65.350	.829	65.726
20	.81 to 1.43	.00	15.12	2.75	.891	52.234	.881	55.376	.881	55.382
20	.81 to 1.43	.25	16.16	2.14	.847	63.802	.832	65.018	.841	63.190
20	.50 to 1.74	.00	14.93	2.79	.872	56.988	.882	55.384	.882	55.470
20	.50 to 1.74	.25	16.36	2.09	.797	71.570	.797	70.720	.804	69.164
40	0.00	.00	31.73	5.55	.940	39.034	.936	40.496	.936	40.496
40	0.00	.25	33.52	4.37	.903	50.188	.898	51.046	.896	50.852
40	.81 to 1.43	.00	31.30	5.53	.935	40.648	.932	41.832	.932	41.848
40	.81 to 1.43	.25	33.47	4.26	.908	49.142	.903	50.554	.905	50.266
40	.50 to 1.74	.00	31.15	5.39	.934	40.788	.939	38.932	.939	38.940
40	.50 to 1.74	.25	33.40	4.16	.890	52.882	.892	52.898	.893	52.678

[1] N = 500

[2] Spearman Rank-Difference Formula

[3] Average absolute difference in rank order

Table 3

Summary of the Goodness-of-Fit Results
(Uniform Ability Distribution,[1] θ = - 2.5 to +2.5)

					Comparison of Estimates					
Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics		True Versus One Parameter Model		True Versus Two Parameter Model		True Versus Three Parameter Model	
			$\bar{X}$	SD	r^2	AAD[3]	r	AAD	r	AAD
20	0.00	.00	9.91	5.84	.970	28.264	.970	28.368	.970	28.368
20	0.00	.25	12.40	4.43	.932	41.850	.931	41.972	.949	36.968
20	.81 to 1.43	.00	9.97	5.63	.969	28.808	.969	29.138	.969	29.140
20	.81 to 1.43	.25	12.28	4.35	.931	42.402	.928	43.932	.943	38.594
20	.50 to 1.74	.00	10.50	5.58	.965	30.826	.966	30.140	.966	30.140
20	.50 to 1.74	.25	12.40	4.54	.932	42.200	.931	42.726	.942	39.016
40	0.00	.00	20.99	12.21	.984	20.438	.984	20.614	.984	20.614
40	0.00	.25	24.54	9.40	.964	30.130	.964	30.260	.971	27.018
40	.81 to 1.43	.00	20.31	12.54	.983	21.088	.983	21.250	.983	21.254
40	.81 to 1.43	.25	24.58	9.36	.962	30.690	.962	30.750	.971	27.738
40	.50 to 1.74	.00	19.93	12.12	.981	22.478	.982	21.814	.982	21.808
40	.50 to 1.74	.25	24.94	9.16	.962	31.490	.964	30.498	.972	27.302

[1] N = 500
[2] Spearman Rank-Difference Formula
[3] Average absolute difference in rank order

Table 4

Summary of the Goodness-of-Fit Results
(Lower Half of Normal Ability Distribution,[1] $\bar{X}_\theta = 0.00$, $SD_\theta = 1.00$)

					Comparison of Estimates					
					True Versus One Parameter Model		True Versus Two Parameter Model		True Versus Three Parameter Model	
Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	SD	r[2]	AAD[3]	r	AAD	r	AAD
20	0.00	.00	6.77	2.69	.817	65.584	.817	65.584	.817	65.584
20	0.00	.25	10.04	2.54	.649	94.928	.649	94.928	.736	82.536
20	.81 to 1.43	.00	6.72	2.66	.835	62.716	.830	63.262	.830	63.312
20	.81 to 1.43	.25	10.10	2.56	.653	95.184	.645	95.774	.729	83.486
20	.50 to 1.74	.00	7.05	2.61	.796	70.646	.801	69.428	.801	69.414
20	.50 to 1.74	.25	10.25	2.57	.655	94.628	.641	95.800	.725	83.380
40	0.00	.00	13.61	5.48	.909	46.026	.909	46.026	.909	46.026
40	0.00	.25	20.06	4.78	.813	68.700	.813	68.700	.848	61.626
40	.81 to 1.43	.00	13.65	5.55	.903	48.234	.908	47.276	.907	47.280
40	.81 to 1.43	.25	20.19	4.86	.810	68.078	.816	67.048	.852	60.094
40	.50 to 1.74	.00	14.29	5.78	.901	48.218	.909	46.580	.909	46.582
40	.50 to 1.74	.25	20.47	4.90	.805	69.010	.813	68.662	.848	61.578

[1] N = 500
[2] Spearman Rank-Difference Formula
[3] Average absolute difference in rank order

Table 5

Summary of the Goodness-of-Fit Results
(Upper Half of Normal Ability Distribution,[1] $\bar{X}_\theta = 0.00$, $SD_\theta = 1.00$)

Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics		Comparison of Estimates					
					True Versus One Parameter Model		True Versus Two Parameter Model		True Versus Three Parameter Model	
			$\bar{X}$	SD	r^2	AAD[3]	r	AAD	r	AAD
20	0.00	.00	13.37	2.62	.844	60.506	.844	60.808	.844	60.808
20	0.00	.25	15.12	2.20	.761	75.752	.759	76.158	.769	75.076
20	.81 to 1.43	.00	13.37	2.61	.853	61.088	.852	61.596	.852	61.606
20	.81 to 1.43	.25	15.12	2.18	.759	76.406	.757	78.024	.769	75.628
20	.50 to 1.74	.00	13.43	2.52	.834	64.792	.846	63.084	.846	63.076
20	.50 to 1.74	.25	15.11	2.12	.749	78.686	.752	79.920	.767	77.012
40	0.00	.00	27.96	4.93	.895	50.714	.895	50.748	.895	50.748
40	0.00	.25	31.02	3.75	.823	65.180	.822	65.448	.833	64.236
40	.81 to 1.43	.00	28.28	4.91	.894	51.252	.898	50.212	.898	50.226
40	.81 to 1.43	.25	31.11	3.81	.824	65.924	.830	64.838	.839	63.160
40	.50 to 1.74	.00	28.39	4.90	.892	51.014	.898	49.954	.898	49.952
40	.50 to 1.74	.25	31.20	3.77	.808	67.604	.822	64.512	.828	63.958

[1] N = 500
[2] Spearman Rank-Difference Formula
[3] Average absolute difference in rank order

Table 6

Summary of the Goodness-of-Fit Results
(Normal Ability Distribution,[1] $\bar{X}_\theta = 0.0$, $SD_\theta = 1.0$)

					Comparison of Estimates					
Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics		True Versus One Parameter Model		True Versus Two Parameter Model		True Versus Three Parameter Model	
			$\bar{X}$	SD	r^2	AAD[3]	r	AAD	r	AAD
20	0.00	.00	10.30	4.27	.940	36.844	.940	36.906	.940	36.906
20	0.00	.25	12.37	3.49	.883	53.940	.883	53.896	.908	47.554
20	.81 to 1.43	.00	10.43	4.33	.943	35.868	.944	35.988	.944	35.982
20	.81 to 1.43	.25	12.40	3.46	.882	54.306	.883	54.336	.905	48.610
20	.50 to 1.74	.00	10.51	4.20	.930	41.114	.932	40.958	.932	40.962
20	.50 to 1.74	.25	12.48	3.50	.873	55.726	.865	57.942	.881	53.128
40	0.00	.00	21.22	9.21	.971	26.598	.971	26.620	.971	26.620
40	0.00	.25	25.78	7.11	.946	36.442	.946	36.464	.956	33.030
40	.81 to 1.43	.00	20.90	9.39	.973	25.196	.973	25.536	.973	25.534
40	.81 to 1.43	.25	25.88	7.01	.939	38.864	.942	37.648	.952	34.148
40	.50 to 1.74	.00	20.87	8.99	.970	27.038	.972	25.878	.972	25.874
40	.50 to 1.74	.25	25.91	6.99	.937	38.794	.941	37.330	.951	34.676

[1] N = 500

[2] Spearman Rank-Difference Formula

[3] Average absolute difference in rank order

distribution (∿.65 to ∿.73). The improvement in the average absolute difference in rank order was about 13.

3. With the forty-item tests, the three-parameter model was also somewhat more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .04 higher in both ability distributions. The improvement in the average absolute difference in rank order was about 8. The reduction in effectiveness of the three-parameter model weights was to be expected with the longer tests. Gulliksen (1950) noted the insignificance of scoring weights when the test gets longer and test items are positively correlated.

4. For examinees in the upper half of the ability distribution, and for the data sets studied, the number rights score was about as effective as the more complicated scoring weights used in the two- and three-parameter models.

Shape of the Ability Distribution

5. As expected, correlations tended to be higher for the uniformly distributed ability scores.

Test Length

6. It is interesting to observe the increases in correlations due to doubling the length of the test. Again, as expected they tended to be rather small.

Conclusions

From the data sets analyzed in this study, it is clear that there are some sizable gains to be expected with modest length tests (n = 20) in the correct ordering of examinees at the lower end of the ability continuum when three-parameter model estimates are used (as opposed to the number right score). The gains were cut roughly in half when the tests were doubled (n = 40) in length. It was also surprising (to us) that item discrimination parameters as weights had so little effect on the results. On the other hand, Gulliksen (1950) had summarized the research on item weights nearly thirty years ago and came to essentially the same conclusion! This brings us to what we feel is a very important point. To the extent that our simulated data sets are typical of real data, it would appear that the application of latent trait models to the problem of "ranking" examinees is probably not worth the trouble except in those situations where gains of the size noted for lower ability examinees in the paper are important. The number right score does nearly as good a job of ranking examinees as the most complicated scoring methods.

We do caution the reader however from generalizing the results from a single study. For one, the authors have not had enough experience fitting the three-parameter model to real data to feel sure about the "typical" values of the item parameters. It is possible that our simulations do not closely reflect real data. Second, our criterion measure of goodness of fit seems suitable for the situation in which a user desires to make norm-referenced interpretations of his/her test scores. There are many other test situations (for example, those involving tailored tests, test score equating, and criterion-referenced tests) where a different criterion to judge the quality of a solution would be more suitable. Third, the results of our study provide a somewhat unfair comparison of the two-parameter model with the other two models. This is because the item discrimination parameters used in the weighting process to derive statistics for ability estimation would have been somewhat different had the "best-fitting" two-parameter curves to the three-parameter item characteristic curves been used. The item discrimination parameters in the "best fitting" two-parameter curves would have differed somewhat from those defined in the three-parameter curves they were fitted to.

A final point should also be stressed. The correlation results of the one-parameter model and (to a much lesser extent) the two-parameter model are inflated (to an unknown extent) because of tied scores. Therefore, the true differences in the reported correlations are somewhat larger than those reported in Tables 1 to 6. This error in our methodology will be corrected before we prepare our paper for publication.

In summary, the future of latent trait theory as a framework for solving educational testing problems has been firmly established. There have already been major breakthroughs in important areas of testing through the use of latent trait theory. It is our hope that our methods and results will encourage others to seek to define and to use other practical criteria for comparing the results of fitting latent trait models to simulated as well as real data to the extent it is possible to do so. Certainly there is substantial need for more research aimed at providing practitioners with practical guidelines for model selection, test design and test score analysis.

References

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

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, 1950.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

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

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D., & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues, and applications. Review of Educational Research, 1978, in press.

Hambleton, R. K., & Traub, R. E. The robustness of the Rasch test model. Laboratory of Psychometric and Evaluative Research Report No. 42. Amherst, MA: University of Massachusetts, 1976.

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.

Lord, F. M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F. M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75.

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

Lord, R. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

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

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Tinsley, H. E. A., & Dawis, R. Test-free person measurement with the Rasch simple logistic model. Applied Psychological Measurement, 1977, 1, 483-487.

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, N.J.: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

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