DOCUMENT RESUME

ED 135 853 TM 006 084

AUTHOR Reckase, Mark D.
TITLE Computerized Achievement Testing Using the Simple Logistic Model.
SPONS AGENCY Missouri Univ., Columbia.
PUB DATE [Apr 77]
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; College Students; *Computer Programs; Higher Education; *Mathematical Models; Predictive Validity; Probability; Statistics; Student Attitudes; *Test Reliability; *Test Validity
IDENTIFIERS *Rasch Simple Logistic Model; *Tailored Testing

ABSTRACT

The reliability and validity of a tailored testing procedure based on the simple logistic model was determined for an achievement test in statistics and measurement. The test was administered on a CRT terminal to students from graduate and undergraduate measurement courses. Equivalent form reliability over a one-week interval was found to be 0.595 while KR20 reliability measures for the traditional course exams yielded 0.756 and 0.574. The tailored testing procedure administered 20 items or less while the traditional tests were 55 and 50 items respectively. The tailored test was a valid predictor for the graduate course. (Author)

Computerized Achievement Testing
Using the Simple Logistic Model

by

Mark D. Reckase

University of Missouri-Columbia

Tailored testing, the selection and administration of items designed to match each person's capabilities, has excited substantial interest since its original development more than two decades ago (Krathwohl & Huyser, 1956). However, the majority of the research in this area has dealt with aptitude tests since they more closely meet the assumptions made by the test models used to implement the procedures. Achievement testing, on the other hand, is an area of enormous potential for a procedure that can be self scheduled, self paced, and which yields, in theory, an unlimited number of equivalent measures that can be used for efficient classroom assessment.

The purpose of this paper is to evaluate the use of tailored testing for the measurement of classroom achievement. To date, minimal research has been done in this area. Reckase (1975) compared the results of testing using traditional methods and simple logistic tailored tests and found the shorter tailored tests to give similar measures to the longer paper-and-pencil forms. A shortcoming of that study was the small sample size, limiting the generalizability of the results.

Ferguson (1969) evaluated an approach for achievement testing based on Wald's sequential analysis. He used 75 children in primary grades for a real data simulation and found the sequential test to yield high reliabilities and high correlation with the parent test. Other than these studies, virtually no research has been published in tailored achievement testing. This paper aims to rectify the lack.

Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977. This research was funded by a grant from the Research Council of the Graduate School of the University of Missouri-Columbia.

The Tailored Testing Model

The tailored testing procedure used in the study reported here has been described in detail elsewhere (Reckase, 1974). However, the major components of the procedure will be briefly summarized to distinguish it from the other techniques currently being used.

The procedure is based on the simple logistic model developed by Rasch (1960). This mathematical model was chosen more for reasons of computational simplicity and speed than any belief that the model precisely described the interaction of examinees and test items. Undoubtedly, the three-parameter logistic or normal ogive models would fit the response data more closely. However, calibration procedures for the three-parameter models are more complex and the on-line ability estimation procedures are more time-consuming, encouraging attempts to apply the simple logistic model.

The simple logistic model is based on the concept that the probability that a person presented with an item will respond correctly is a function of two parameters: the ability of the person and the easiness of the item. The basic formula of the model is given in Equation 1 (Rasch, 1966).

$$P\{X_{si}\} = \frac{(A_s E_i)^{X_{si}}}{1 + A_s E_i}, \quad X_{si} = 0,1, \qquad [1]$$

where A_s is an ability parameter for Person s, E_i is an easiness parameter for Item i, and X_{si} is the item score.

The simple logistic model is a special case of Birnbaum's (1968) three parameter logistic model.

$$P\{X_{si} = 1\} = c_i + \frac{(1 - c_i)e^{a_i(\theta_s - b_i)}}{1 + e^{a_i(\theta_s - b_i)}} \qquad [2]$$

where c_i is a guessing parameter, a_i is a discrimination parameter, and b_i is a difficulty parameter for Item i, and θ_s is an ability parameter for Person s. Equation 1 can be derived from Equation 2 by setting $c_i = 0$, $a_i = 1$, $e^{\theta_s} = A_s$, and $e^{-b_i} = E_i$. Excellent presentations of the properties of the simple logistic model have been given by Rasch (1960) and Whitely and Dawis (1974). Birnbaum (1968) has described the three-parameter model in detail.

In order to use the simple logistic model in a tailored testing environment, a pool of items is first calibrated to obtain estimates of the easiness parameters. These parameters are stored with the items in the computer and are used in item selection. The calibration of items for the study was done using a modified version of a program given in an article by Wright and Panchapakesan (1969).

The actual testing procedure used has been classified by Weiss (1974), as a variable branching, maximum likelihood procedure. The procedure begins testing with an item of median difficulty for the examinee's estimated ability level if that information is available or an item of median difficulty for an average person if there are no previous estimates. A fixed stepsize up-and-down procedure is then followed until both a correct and incorrect response are obtained. At that point, the ability parameter is estimated and the next item is selected to have 0.5 probability or larger of a correct response for the estimated ability. However, if no items within ±0.3 of this value are available, the session is terminated. After each item is administered, a new ability estimate is computed and it is used to select the next item. Administration of items continues until all appropriate items in the pool are used, ability has been estimated to sufficient accuracy, or a set number of items have been administered.

Research Design

The purpose of this study was to evaluate the reliability and validity of a test administered by the simple-logistic tailored testing procedure. The procedure was evaluated as a device for estimating and predicting academic performance on a statistics and measurement unit in courses at the University of Missouri-Columbia. The study took place during the Summers of 1975 and 1976.

The sample used for the study was composed of 73 students from graduate and undergraduate measurement courses at the University of Missouri who volunteered to participate in the experiment. The students ranged from college juniors to second year graduate students. Each student was examined twice using the tailored test and was administered an attitude scale concerning the testing procedure. Information about course achievement was also available for use as a criterion for validation of the tailored achievement test. The experiment was conducted as follows.

During the week immediately following an examination of statistics and measurement in two measurement courses, students were asked to sign up for two testing sessions exactly one week apart. On the date of each student's testing session, each was reminded of his appointment to reduce the number of "no shows." At the session, each student was examined using a cathode ray terminal that had been connected to the IBM 370/168 computer at the University while they signed in for the experiment.

During the testing session, each student was administered items until those within ±0.3 log easiness of that required by the program were no longer available or until twenty items had been administered, whichever came first. The ±0.3 log easiness rule resulted from research which showed that using items greatly deviating from those requested by the program induced bias in the maximum likelihood routine (Reckase, 1975). The twenty item limit was imposed to keep the testing time within thirty minutes.

The item pool for the tailored testing procedure was made up of 96 statistics and measurement items that had been stored in the computer with calibration data obtained using the Wright and Panchapakesan (1969) program mentioned earlier. From 250 to 966 students were administered the various test items to obtain the data for the calibration program.

Exactly one week following the first testing, each student was tested again using the procedure given above. However, the entry point into the item pool on the second testing was based on each student's ability estimate from the first testing, rather than at the central point used initially. Thus each student received quite a different set of items on the second testing. After the second testing session a short attitude questionnaire was administered to determine time pressures, perceived difficulty, anxiety, and procedure preference.

From the two testing sessions and achievement measurements from the course, the following data were gathered on each student: (a) the responses to the four item attitude questionnaire; (b) the ability estimate from the first testing session; (c) the number of items administered at the first session; (d) the ability estimate from the second session; (e) the number of items administered at the second session; (f) the statistics and measurement exam scores; (g) the final exam scores; and (h) the number of days between the statistics and measurement exam and the first testing session.

The information collected was then analyzed to answer three questions: (a) How reliable is the achievement test administered using tailored testing as compared to the paper-and-pencil class exams; (b) Is the tailored test or the traditional test better for predicting the final exam scores; and (c) What attitudes do the students have toward the tailored testing procedure?

Results

The descriptive statistics summarizing the results are given in Tables 1 and 2. Table 1 gives the mean, median, standard deviation, skewness, and kurtosis for each of the variables measured in the study for the graduate and undergraduate groups. Since no evidence was found to indicate significant differences in the mean values of the variables for the two groups, summary data for the combined group is given at the bottom of Table 1. The separate groups are maintained despite their similarity because the Exam II and Final Exams are different for the two groups requiring separate correlations in the validation of the tailored tests.

Insert Table 1 about here

Based on the skewness and kurtosis values, the distributions for the first tailored testing are not significantly different from the normal distribution, while the second testing yields scores with a slight negative skew. No difference was found between the mean score on the first testing and that obtained on the tailored test a week later. The distributions for the number of days after the class exam until the first tailored testing, the number of items on the second tailored test are significantly platykurtic, indicating the presence of a substantial number of both high and low values for these variables.

An important summary statistic for this study is the number of items administered in the tailored testing sessions. The mean number of items for the first session was found to be 13.68 and for the second session was 12.08. The reduction is significant $t(58) = 1.98$, $p < 0.05$ indicating the effect of changing the starting point in the item pool. Since the paper-and-pencil tests were 50 items long, these results indicate approximately a 75% reduction in the number of test items used to measure the student's ability.

Table 2 gives the values of the correlations between the relevant variables for the graduate (below the diagonal) and undergraduate (above the diagonal) groups. Since many of the correlations were found to be significantly different for the two groups, the correlations were not determined for the combined group.

Insert Table 2 about here

The reliability of the tailored test is obtained from the correlations between the first and second tailored testing session. The reliability technique used here is equivalent forms over time, since only about a third of the items were the same as on the first testing. The values obtained (0.49 for the undergraduate group and 0.69 for the graduate group) correspond favorably to the KR-20 values of 0.53 and 0.76 obtained for the traditional tests on the same material for the undergraduate and graduate groups respectively. These values are especially close considering the conservative nature of the equivalent-forms-over-time reliability method.

The correlation of the tailored test with the second course exams also gives evidence for the reliability of the procedure since the correlations with the paper-and-pencil tests are about the same as the test-retest values. Thus the paper-and-pencil tests yield forms that are as parallel as the retest with the tailored test.

The validity of the tailored test for predicting an outside criterion can be estimated from the correlations with the final exam scores for the two groups. The final exams cover substantially different material than the tailored test or the paper-and-pencil second exam, making it an appropriate test for predictive validity. For the undergraduate class, the paper-and-pencil test had a substantially higher correlation than the tailored test with the final exam scores (0.82 vs. 0.54) while for the graduate group the correlations were about equal for the two tests.

The correlations between the responses to the attitude items and the other variables generally show that the attitude items tend to correlate with each other, but not with the test results. The only significant correlations are between the item concerning perceived test difficulty, the first tailored test results, and the number of items on that test. In effect these correlations indicate that students who attained high scores on the tailored test found it easier and took fewer items.

The attitude items and response data are presented in Table 3. The items were tested against a rectangular distribution of responses using a χ^2 test to determine if there were any significant response tendencies in the data. The results show that students felt equal time pressure on the tailored test and paper-and-pencil test, and found the tailored tests harder. No significant preference was found between the tailored test and paper-and-pencil test and the two testing settings were evenly divided in their anxiety producing effects.

Insert Table 3 about here

Discussion

The purpose of the research reported in this paper was to try out a simple-logistic tailored testing procedure for use in classroom achievement testing. Two fairly different types of students took part, giving some evidence for the generalizability of the findings. Also, the simple-logistic model is identical to the equal discrimination, no guessing case of the three parameter logistic model which has been used in several simulation studies (McBride, 1976; Lord, 1970). Thus the results presented here should begin to fill in gaps in currently available research results on tailored testing.

The results of the study can be organized into four areas: (a) reliability; (b) efficiency; (c) validity; and (d) attitude. The reliability data generally show that the tailored test yields essentially an equivalent form to the traditional paper-and-pencil tests. This is indicated by the similarity of the correlation of the tailored test with the paper-and-pencil test to the paper-and-pencil test reliability. The equivalent forms over time reliability is also consistent with this interpretation. The reliabilities found in this study are not extremely high, but they fall in the range typical of classroom tests. The tests used also tended to be multidimensional reducing the possible values of the internal consistency measures, and degrading the unidimensional parameters of the logistic model.

The data on efficiency relate to the number of items required by the testing procedure and the amount of time required for the testing session. On the average, about twelve items were required for the tailored test compared to fifty on the paper-and-pencil test showing a substantial saving. Also, the testing sessions averaged about a half hour for the tailored test while fifty minutes were required for the paper-and-pencil test. Thus both in terms of time and items used the tailored test reduces the requirements without losing reliability.

The data on predictive validity reported in the study did not yield definitive results. One group (graduate) yielded validity coefficients about equal to the paper-and-pencil test for the tailored test, while the other group yielded coefficients substantially lower. This fact may be explained by a difference in motivation for the students taking the tailored test since it did not count toward their grade while the paper-and pencil test did. Method variance may also explain some of the differemces since the final exams were in the traditional format. Further research using the tailored testing procedure for course exams should determine the importance of motivational effects.

The attitude items showed little relation to performance on the tests, possibly a result of motivational factors. The students showed neither more nor less anxiety or preference for either procedure, indicating that the use of the terminal does not cause fear or uncertainty in the student. The tailored tests did seem hard to the students. This seems reasonable since the brighter students did receive much harder items than they were used to on exams.

In summary, the results show the tailored testing procedure yields results that are as good as traditional tests in measuring achievement with a substantial reduction in test length and testing time. The predictive validity may be somewhat less than traditional tests, but this area required further study before a definitive conclusion can be made. Students preferred the tailored testing procedure equally to the traditional test and expressed no increase in anxiety when taking the test on the terminal. However, the tailored tests did seem somewhat harder. Overall, the simple-logistic tailored testing procedure seems to be a reasonable alternative to the traditional classroom achievement test.

Table 1

Descriptive Statistics

Group	Statistics	Number of Days	First Tailored Test	Number of Items	Second Tailored Test	Number of Items	Exam II[a]
Undergrad (N = 43)	Mean	4.23	1.87	13.71	2.02	12.06	51.16
	Median	5.06	2.17	13.88	2.23	12.00	51.00
	Standard Deviation	2.53	1.35	5.72	1.16	7.15	9.94
	Skewness	-0.51	-0.34	-0.05	-0.41	-0.10	-0.22
	Kurtosis	-1.28	0.11	-1.73**	-0.76	-1.69**	-0.87
	N	43	41	41	35	35	43
Graduate (N = 30)	Mean	2.63	1.94	13.63	1.87	12.11	54.13
	Median	2.25	2.25	15.00	1.90	12.50	53.50
	Standard Deviation	2.17	1.44	6.32	1.52	6.49	10.65
	Skewness	0.66	-0.08	-0.11	-0.90	-0.26	-0.08
	Kurtosis	-0.38	-1.05	-1.87	-0.49	-1.34	-0.93
	N	30	27	27	28	28	30
Total (N = 73)	Mean	3.58	1.90	13.68	1.95	12.08*	
	Median	3.95	2.24	14.00	2.22	12.38	
	Standard Deviation	2.50	1.38	5.92	1.33	6.81	
	Skewness	-0.02	-0.22	-0.08	-0.52	-0.61	
	Kurtosis	-1.42**	-0.36	-1.76**	-0.30	-1.53**	
	N	73	68	68	63	63	

[a] The Exam II and the Final were different for the two groups and, therefore, statistics have not been combined for the total group.

*p ≤ 0.05
**p ≤ 0.01

Table 2
Correlations Between Variables by Group
(Undergrad above diagonal, grad below diagonal)

	Variable Number										
Variable	1	2	3	4	5	6	7	8	9	10	11
1. First Tailored Test	--	-79**	49**	59*	49**	54*	-08	-11	13	-01	-08
2. Number of Items	-88*	--	-37*	52**	-30*	-48*	-16	-04	-21	-01	08
3. Second Tailored Test	69**	-54**	--	-85**	45*	35	-05	-08	14	-11	03
4. Number of Items	-55**	51**	-80**	--	-41**	-36	00	18	-12	08	-02
5. Exam II	65**	-57**	66**	-53**	--	82*	-01	-05	16	-14	04
6. Final Exam	27	-17	62*	-59*	69*	--	-24	-07	-08	-22	-20
7. Number of Days	01	-16	-17	-02	-18	-08	--	13	-05	-10	-13
8. Question 1	-20	26	-25	15	-15	-37	-13	--	19	37*	43*
9. Question 2	37*	-48**	28	-30	02	-14	11	01	--	42**	43*
10. Question 3	07	-22	08	-31	04	-16	-05	41*	48*	--	49*
11. Question 4	-07	02	05	-20	-12	-55*	-12	52**	38*	57**	--

*p ≤ 0.05
**p ≤ 0.01

Tabl

Attitude Item ata

1. Compared to multiple choice tests, the tailored test had

	Response Frequency*	Value[a]
(a) more time pressure.	12	1
(b) less time pressure.	17	3
(c) about equal time pressure.	35	2

2. Compared to traditional multiple choice tests, the tailored test is

	Response Frequency*	Value
(a) easier.	15	3
(b) harder.	45	1
(c) about as difficult.	4	2

3. As compared to the traditional multiple choice test,

	Response Frequency	Value
(a) I would rather take the tailored test.	27	3
(b) I would rather take the traditional test.	16	1
(c) I prefer both equally well.	21	2

4. Taking the test on the computer makes me

	Response Frequency	Value
(a) more anxious than a traditional test.	19	1
(b) less anxious than a traditional test.	29	3
(c) about equally as anxious as the traditional test.	16	2

*Response distribution is significantly different from a rectangular distribution $p \leq 0.05$.

[a] These are the values assigned to the responses for use when correlating the items with the other variables.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968, Chapters 17-20.

Ferguson, R. L. The development, implementation, and evaluation of a computer-assisted branched test for a program individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Green, B. F. Presentation at the Invitational Conference on Adaptive Testing, Washington, D. C., June, 1975.

Krathwohl, D. R. & Huyser, R. J. The sequential item test (SIT). American Psychologist, 1956, 2, 419.

Larkin, K. C. & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing. (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, June, 1974.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing and guidance, New York: Harper & Row, 1970.

Lord, F. M. Presentation at the Invitational Conference on Adaptive Testing, Washington, D.C., June, 1975.

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

Rasch, G. An individualistic approach to item analysis. In P. W. Lazarsfeld and H. W. Henry (Eds.) Readings in mathematical social science. Chicago: Science Research Associates, 1966.

Reckase, M. D. An interactive computer program for tailored testing based on the one parameter logistic model. Behavior Research Methods and Instrumentation, March, 1974, 6 (2), 208-212.

Reckase, M. D. The effect of item choice on ability estimation when using a simple logistic model. Columbia, Mo.: University of Missouri, 1975. (ERIC Document Reproduction Service No. ED 106 342).

Siegel, S. Non parametric statistics, New York, McGraw-Hill, 1956.

Weiss, D. J. *Strategies of computerized ability testing*, (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, 1974.

Whitely, S. E. & Dawis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, Fall, 1974 II (3), 163-178.

Wood, R. Response-contingent testing. *Review of Educational Research*, Fall, 1973, 43 (4), 529-

Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, 29 23-48.