DOCUMENT RESUME

ED 166 218 TM 008 021

AUTHOR Slawski, Edward J.; Bauer, Ernest A.
TITLE Reducing Testing Time While Preserving Test Information: A Ten Item Fourth Grade MEAP Reading Test.
PUB DATE 17 Mar 78
NOTE 32p.; Paper presented at the Annual Meeting of the Michigan Educational Research Association (Detroit, Michigan, March 17, 1978) ; Table 2 may reproduce poorly due to small type

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Objectives; Grade 4; Intermediate Grades; *Item Analysis; *Mastery Tests; Reading Achievement; *Reading Tests; Scores; Scoring; *Statistical Analysis; Student Testing; *Test Interpretation
IDENTIFIERS Michigan; Michigan Educational Assessment Program; Rasch Model; *Test Length

ABSTRACT

A new method of analysis was used in the Michigan Educational Assessment Program to test minimum competencies in fourth grade reading achievement. This technique permitted a substantial decrease in testing time and costs. The original test consisted of 95 items measuring 19 objectives; mastery was indicated by correct responses to four out of the five items measuring each objective. Data from those 1,096 students whose raw scores were between 36 and 83 were re-analyzed. Several tests were used to determine which items were acceptable for analysis using the Rasch model. It was felt that the 95 items fit the Rasch model, and that the item calibrations would yield standard log achievement scores (SLAs) that would accurately summarize where students fell on the latent trait measured by the test. These SLAs provided essentially the same information as the number of objectives mastered, with a shorter test. For most students, the items were relatively simple given student achievement levels; therefore, the amount of information provided was slight. A short, ten-item test was developed; analysis indicated that it imposed a more rigid criterion than the longer test. Mastery decisions using the short test led to 14.9% fewer Type II errors (false negatives) and 1.6% more Type I errors (false positives). (GDC)

Reducing Testing Time While Preserving Test Information

A Ten Item Fourth Grade MEAP Reading Test

Paper Presented at Annual Meeting of Michigan Educational Research Association, Detroit, Michigan - March 17, 1978

Edward J. Slawski
Pontiac Public Schools

Ernest A. Bauer
Oakland Schools

Introduction

For too long, many in the educational community have ignored the information that has been available to help make informed educational decisions. It was not uncommon for the results of the administration of batteries of tests to gather dust in offices and never be consulted before educational plans were formulated. Recent pressures for improving the delivery of basic skills in public education, occurring as they do at a time when the amount of monies available for public education seems to be decreasing, have intensified the needs of educational decision-makers for more effective educational plans. Test results provide one important source for information to make better educational plans. However, even though the use of such data is increasing, the administration of tests and the analysis of results is expensive in terms of instructional time and dollar costs. This paper will present an introduction to a new method of test analysis and its application in creating an alternative to current testing practice in State Assessment. This alternative seems to permit a substantial decrease in total testing time and costs without substantial loss in the information provided.

The Michigan Educational Assessment Program

For the past nine years the Michigan Educational Assessment Program (MEAP) has endeavored "to provide information on the status and progress of Michigan basic skills education" to state and local educational decision-makers and their clients.[1] The assessment is carried out by administering objective-referenced tests in "an important, but limited number of minimal skills in reading and mathematics" at grades four and seven throughout the state.[2] The results of these tests "provide for standard measurement of all pupils and help control personal bias and arbitrary judgments by educational decision-makers." In addition, test results provide input "when curricula (sic) decisions are being made by curriculum specialists, both at the state and local levels." Moreover, MEAP test results are used "to identify high needs schools" so that the state can "initiate contacts with local school districts and offer to help them in addressing the achievement problems there." More generally, "it is considered appropriate for the state to use MEAP test results as part of the process for allocating state funds." Since so many important educational programs and individual student decisions are based upon the results of these tests, a full understanding of the composition and performance of these tests is crucial.

In order to enable users to understand the characteristics of the MEAP tests, the Research, Evaluation and Assessment Services of the Michigan State Department of Education (MDE) publishes a comprehensive Tehcnical Report (MEAP, 1976) which includes various item and objective statistics.

---

[1]The quotations that appear in this paragraph are taken from a pamphlet published by MEAP, entitled "Do YOU Use MEAP Tests Appropriately?" The pamphlet is distributed to local district users of the tests results to assist in the appropriate use and interpretation of MEAP test data.

[2]The State has been piloting experimental versions of first and tenth grade MEAP tests.

Validity, reliability and item discrimination measures are provided. These data indicate that the tests perform acceptably as tests.[1] The purpose of this paper will be to report an examination of one of the tests, Grade Four MEAP Reading, under the assumptions of the Rasch model. We report these results not to be critical of the current procedures used by the Michigan Educational Assessment Program but to explore additional procedures and techniques of analyzing the tests and reporting the result of statewide assessment.[2]

The Grade Four MEAP Reading Test was chosen for the present analysis. The test consists of 95 items which measure 19 fourth grade reading objectives. There are five items for each objective on the test: mastery is reached if the student correctly answers 4 of 5 items.[3] According to the State Summary, 61 percent of the pupils mastered 75 percent of the objectives statewide. No item data were reported in the summary.

The MEAP state sample tape provided a random sample of the results of the Grade Four Reading Test for approximately 5000 fourth grade students in Michigan. A random half of these students were selected for the present analysis, yielding a case base of 2568 subjects.[4] For the students in this analysis, the mean number of objectives mastered was 13.6 and the median was 16.2 objectives attained. The mean number of items answered correctly by this sample group was 74.8 and the median was 83.0 items.

---

[1] In a series of four articles, Rudman (1977a,b,c,d) has offered some criticism of the tests based upon his analysis of the traditional measures of test statistics. Our purpose here prevents us from exploring either his or other criticism of MEAP tests.

[2] We wish to thank Research, Evaluation and Assessment Services of the Michigan Department of Education for making available the data which supported this analysis.

[3] The latest available Technical Report (MEAP, 1976:18-25) reports reliability and item discrimination measures for each objective considered as a five item test. These data indicate that eleven (11) of the objectives have KR-20 Reliability Coefficients greater than .70 and one is below .49. The median phi coefficient for the association between objective and item attainment was .88.

[4] The SAMPLE procedure provided by SPSS was used to select a random half (Nie, and others, 1975: 127-8).

Rasch Models

Recent developments in latent trait theory have occasioned a renewed interest in "true score theory" (Lord and Novick, 1968). Under the leadership of Wright and his students (Wright, 1968, 1977; Wright and Panchapakesan, 1969; Wright and Mead, 1977), a latent trait model originally proposed by Georg Rasch (1960, 1966) has caught the attention of the educational measurement community (see, for example, Journal of Educational Measurement, 14 (Summer), 1977). Under the assumptions of the Rasch model, an individual's score is "governed by the product of the ability [achievement level][1] of the person and the easiness of the item" (Wright, 1968:4). The equation which specifies the relationship between the achievement level of the subject and the difficulty of the item can be written:

$$P_{vi} = \frac{e^{(AL_v - D_i)}}{1 + e^{(AL_v - D_i)}}$$

where P_{vi} is the probability the person v correctly answers item i, AL_v is the achievement level of person v, and D_i is the estimate of the difficulty of item i. AL is the Rasch standard achievement score expressed in log achievement units and D represents the Rasch log item difficulty score. These parameters are estimated from the distribution of raw scores and the P values of the items comprising the test.[2] The result of fitting a set

---

[1] Rentz and Bashaw (1977:161) note that reference to "ability" sometimes causes confusion which is unnecessary "if one is aware that 'ability' as used here is a generic term that means the trait or characteristic of the examinee being measured by the particular test under consideration."

[2] Birnbaum (1968: 402) notes that the Rasch model "is a special case of the logistic model in which all items have the same discriminating powers, and all items can vary only in their difficulties." Hambleton and Traub (1970) demonstrated that some information is lost by not fitting additional parameters. However, they note that a considerable increase in cost and clarity is incurred by fitting additional parameters.

of items or persons to this model is an interval measure of achievement and item difficulty in terms of the same units. This facilitates examinations of test items, student performance on tests, and instructional content that were impossible under traditional measurement techniques.[1] These features result in "person-free" and "test-free" measurement.[2]

---

[1]Tinsely and Dawis (1972) demonstrate that decisions about items do not differ markedly under Rasch techniques and traditional techniques for choosing test items. The case is that Rasch techniques allow for the selection of items as efficiently as traditional techniques in addition to providing additional measurement power.

[2]There is some dispute with regard to the extent to which measurement is "person free" and/or "test free." However, the model has been found to be relatively robust under violation of assumptions given large enough sample sizes and "fitting" items (Tinsely and Dawis, 1972).

Test Construction

An initial calibration of the 95 items[1] which constitute the Grade Four MEAP Reading Test was performed including only those students whose raw score was between 36 and 83. These score boundaries were chosen with reference to the chance of a student correctly answering an item by guessing if there are four choices for each item. The score interval represents one and a half the chance level for students at the low end of the distrubution (1.5 x 24) and one half the chance level at the top end (95 - (.5 x 24)).[2]

In the initial calibration, 1438 subjects were excluded from the calibration because their scores were outside the specified range: 34 students were immediately excluded because they achieved perfect scores; 224 students received a raw score below 36; 1214 students scored above 83. The first calibration was performed on 1096 students. The mean number of items correctly asnwered by this group was 54.8 and the median was 59.2. The mean standard

---

[1] One of the assumptions of the Rasch model is that the item's are drawn from a homogeneous domain of content. Although the model seems to be robust under violations of this assumption, a factor analysis of the 95 items on the Grade Four MEAP Reading Test was performed to test the dimensionality of the item set. This analysis yielded only one factor with an eigenvalue greater than one and explained 61 percent of the variance among the 95 items. This seems to be reason to believe that the 95 items lie along a single dimension.

[2] Cypress (1973: 4) found that "the estimates derived from the Rasch Measurement Model were not independent of the group used to produce them. Differences were minimal in the middle score range, but large in high and low score ranges." More stable estimates would seem obtainable from subjects in the middle range of scores, those within the boundaries which we established.

Justification for the choice of these boundaries rests in our intuitive unwillingness to believe that low scoring students are "informed guessers." Moreover, the middle range of scores seems to provide more stable information about item difficulty estimates. Robert Rentz, in a personal communication with the authors, stated that our procedures are perhaps more rigorous than necessary and that we may be too willing to believe the tests of fit provided by the model. Rentz prefers to calibrate on all persons. One of the perplexing aspects of work with the Rasch model is the unavailability of any good decision rules for procedural issues.

achievement estimate for the students in the calibration was 219.5 with a standard deviation of 14.8.[1] The results of this calibration were examined to determine how well this set of items "fit" the model.

There is no single statistic which measures the fit of a set of items to the Rasch model. Therefore, we used a series of "tests" to determine whether the 95 items for fourth grade reading performed acceptably. First, we examined the Total Fit Mean Square (FMS) which is computed from each of the 95 items. This statistic represents the mean squared standard residual between how an individual person of a given achievement level performed on items and how he/she could be expected to perform given the difficulty of the item, averaged over persons. Wright and Mead (1977: 50) suggest that this statistic "will be large for an item if there are too many high ability persons who failed on an item and/or too many low ability persons who succeeded." These values averaged over items yield a summary "fit statistic." The value for this statistic obtained from the initial calibration of 95 items was .97 with a standard error of .166. We know that a standard error as high as .20 has been obtained in simulated data that fit the model and so a value of .166 does not seem "too large."

Another indicator of test fit is the ratio between the observed standard error of Total FMS and the one expected given the assumptions of the model over the particular set of items on which the calibrations were done. The expected FMS in these data was .043. Our procedures include the computation of the ratio between the observed standard error of FMS and the expected standard error - the value of which in this case was 3.88. Again, although there are no "rules" for assessing the magnitude of this number, experience indicates that a value of 3.00 or less is desireable. Therefore, the ratio

---

[1]The original Rasch achievement scores are in log units with a mean of 0 and a standard deviation of 1. We have followed standard practice and transformed these scores to a distribution with a mean of 200 and a standard deviation of 10.

of observed to expected FMS is more elevated than we would like it to be before we are willing to believe that the items fit the model.

Finally, the BICAL program (Wright and Mead, 1977) routinely computes the item characteristic curve for each of six different score groups, ranging from extremely low scorers to extremely high scorers. How well the individuals in each of these different score groups perform on the items is measured by a Group Mean Square (GMS) and its standard deviation. The standard deviations may be treated as an χ^2 with one degree of freedom (Wright and Mead, 1977: 37-39). Table 1 displays the GMS and standard deviations for each of the separate score groups. The critical value for χ^2 with 1 df at .01 is 6.6. Therefore, from Table 1, we see that the subjects in the lowest and the highest score group differ significantly in their performance with respect to the model.[1] The distributions of the item statistics that were produced by the initial calibration are displayed in Table 2.

TABLE 1 ABOUT HERE

TABLE 2 ABOUT HERE

[1]We do not want to make too strong a claim about the exact distribution of these numbers. However, the values of the standard deviations in the extreme groups look sufficiently different from the values in the middle four groups for us to wonder about how well the items fit.

TABLE 1

Mean Squares and Standard Deviations for Six Score Groups an initial calibrations of 94 item test

Score Range	36-53	54-64	65-72	73-77	78-80	81-83
Mean Achievement Level	198.8	205.1	211.1	215.3	218.2	220.9
Group Mean Square	9.5	2.6	1.6	2.5	3.1	5.9
SD (GMS)	11.3	3.4	2.1	2.6	5.1	8.8
(Number)	(184)	(188)	(197)	(164)	(151)	(212)

TABLE 2

Item Fit Statistics of 95 Items of

Grade Four MEAP Reading Test

SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ
1	145	-1.42	1.53	0.58
2	152	0.14	1.29	0.90
3	173	0.99	1.08	0.98
4	181	-0.04	1.56	0.79
5	192	0.33	1.42	0.86
6	183	-0.36	1.19	0.93
7	184	0.36	1.36	0.89
8	185	-0.81	1.57	0.66
9	186	0.27	1.42	0.86
10	197	0.37	1.13	0.95
11	165	-0.27	1.24	0.89
12	166	0.16	0.98	1.01
13	167	-0.27	1.16	0.94
14	168	0.71	1.00	1.01
15	169	0.37	1.01	1.00
16	116	-2.18	0.55	1.36
17	117	-0.65	0.90	1.19
18	118	-1.06	1.02	0.97
19	119	-2.22	1.21	0.78
20	120	-1.49	0.99	0.98
21	106	-0.38	0.51	1.19
22	107	0.29	0.35	1.22
23	108	0.59	0.17	1.25
24	109	0.19	0.32	1.27
25	110	0.10	0.26	1.27
26	127	-0.42	0.97	0.99
27	128	-0.83	1.15	0.93
28	129	-1.45	1.36	0.79
29	130	-1.53	1.10	0.94
30	131	-1.29	1.14	0.93
31	135	-0.46	1.41	0.79
32	136	-0.53	1.42	0.76
33	137	-0.63	1.39	0.78
34	138	-0.36	1.50	0.76
35	139	-0.71	1.46	0.73
36	124	0.09	0.99	1.00
37	132	0.01	1.11	0.96
38	133	-0.56	1.30	0.84
39	176	-0.41	1.33	0.85
40	198	2.08	0.59	1.22
41	141	-0.74	1.28	0.83
42	153	-0.17	0.85	1.05
43	174	0.05	1.32	0.87
44	189	1.22	1.22	0.87
45	197	1.96	0.74	1.10
46	121	0.37	0.56	1.14
47	140	0.91	0.30	1.20
48	151	0.19	0.97	1.02
49	170	0.76	0.27	1.23
50	196	0.49	0.98	1.01

SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ
51	134	1.14	0.81	1.07
52	143	0.38	0.69	1.11
53	180	0.36	1.26	0.92
54	190	1.85	0.58	1.18
55	199	1.44	0.90	1.06
56	142	0.33	1.02	0.99
57	148	-0.43	1.13	0.95
58	172	1.12	0.60	1.11
59	177	0.96	1.14	0.97
60	188	2.19	0.42	1.25
61	147	-1.10	1.63	0.63
62	149	-0.87	1.44	0.73
63	175	0.15	1.27	0.91
64	179	0.70	1.31	0.92
65	193	0.76	1.05	1.00
66	111	-0.24	0.70	1.13
67	112	-1.11	0.98	0.99
68	113	0.08	0.68	1.12
69	114	-1.77	0.80	1.15
70	115	-0.84	1.29	0.87
71	123	0.29	1.05	0.99
72	144	-1.06	1.32	0.76
73	150	-0.34	1.14	0.97
74	191	0.15	1.39	0.86
75	1100	0.81	0.95	1.03
76	122	0.74	0.39	1.20
77	146	0.15	0.70	1.11
78	171	1.17	0.60	1.13
79	182	0.79	1.10	0.98
80	195	0.81	1.29	0.92
81	155	-0.92	1.49	0.72
82	156	0.37	1.05	1.00
83	157	-1.53	1.55	0.60
84	158	0.75	1.18	0.97
85	159	-0.53	1.10	0.97
86	160	-0.08	1.30	0.88
87	161	-0.74	1.37	0.81
88	162	1.31	0.46	1.16
89	163	0.23	1.22	0.93
90	164	-0.67	1.20	0.88
91	125	0.10	0.84	1.04
92	126	-0.88	1.12	0.88
93	154	0.81	0.66	1.09
94	178	0.80	1.08	0.98
95	194	1.27	0.73	1.11

Rasch Scores and Number Of Objectives Mastered

In the preceeding section we explored the extent to which the 95 items that comprise the Grade Four MEAP Reading Test "fit" the Rasch model. Our conclusion was that the items fit reasonably well and that the calibrations of the items would yield standard log achievement scores (SLAS) that would accurately summarize where students fall on the latent trait measured by the fourth grade reading achievement test. In this section we will explore how these SLAS are related to other summary measures of student achievement that are currently reported from the results of MEAP testing.[1] We will attempt to show that SLAS provide essentially the same information as number of objectives mastered. In the following section, however, we will demonstrate that SLAS allow for the creation of insturments which can provide for a substantial saving in testing without a loss of information.

One summary measure which enjoys wide use (despite the disclaimers of educators responsible for MEAP) is the proportion of students who master 75 percent of the 19 reading objectives. Many see this statistic as an overall picture of the general level of reading. If a sufficient number of students master 75 percent of the objectives, a reading program is thought to be doing an adequate job of delivering "minimal skills." If the proportion of students mastering 15 objectives falls below a certain level, the district may qualify for additional funds to support improving the delivery of those "minimal skills." In the face of opposition to the use of such measures

---

[1]We realize that there is important information about the performance of students on discrete reading objectives which is not captured in any summary statistic and that this information is important in making instructional decisions at the district, building and student level. We do not argue that summary measures can replace such data. However, another analysis by the authors (in preparation) will examine the utility of Rasch scores to reproduce the information contained in the mastery of discrete objectives and indicate ways in which tests can be redesigned to improve the quality of information about students' achievement with reference to discrete reading skills.

derived from objective-referenced tests, single number summaries are provided and are used to support educational policy decisions. It seems reasonable, then, to compare the performance of SLAS to the number of objectives mastered in order to determine if the Rasch-derived scored provided at least as much information as number of objectives mastered. Any new summary measure ought to work at least as well as the one it replaces.

There is a high positive correlation between SLAS and number of objectives mastered ($r = .93$). Decisions tend not to be based upon the entire range of the numbers of objectives mastered but to be concentrated at that point which seems intuitively to indicate mastery in a more global sense, that is, at 75 percent. Therefore, one way to examine the relationship between SLAS and number of objectives mastered is to establish a criterion level for SLAS which is comparable to mastering 15 reading objectives. Two considerations guided our selection of a SLAS criterion score. First, we noted that MEAP defines mastery of each objective at four correct of the five items which comprise the objectives, that is, 80 percent. Second, in other applications of the Rasch technique to criterion-referenced tests (Kifer and Bramble, 1974), the SLAS which corresponded to correctly answering 80 percent of the items on the test was applied. Therefore, we chose to set the SLAS criterion score at 216, the score which students who answered 76 items correctly received. The question we now examine is whether we would make the same mastery decisions about students using a SLAS criterion score of 216 as we would using mastery of 75 percent of the 19 reading objective at grade four.

Students in the sample were coded into two groups: those who mastered 15 or more objectives and those who mastered 14 or fewer. A distribution of SLAS was prepared for each group. These distributions appear in Table 3. We see that the SLAS distributions are considerably different between masters and non-masters. The median SLAS for those students who mastered 14

or fewer objectives is 206 as compared to a median of 229 for those who

TABLE 3 ABOUT HERE

mastered 15 or more reading objectives. Clearly the distributions are different and the SLAS criterion score seems to sort students into mastery groups that have a similar composition to groups selected on the basis of mastering 75 percent of the objectives. The data summarized in Table 4 present the similarities more explicity. The cross tabulation of the two criteria for mastery shows that in the overwhelming majority (94.9 percent)

TABLE 4 ABOUT HERE

of cases, each criterion yields the same decision about the mastery level of the student. Over one third (34.1 percent) of the sample fail to master 15 objectives and score below 216; three-fifths (60.8 percent) master at least 15 objectives and score 216 or higher. For about one student in twenty (4.8 percent), however, a score of 216 or higher is obtained even though they do not master at least 15 objectives. We suspect that these students either consistently master three of the five items in the objectives or master five of five for a limited number of objectives. In either case, their SLAS will be higher because of the relationship between SLAS and raw score dictated by the Rasch model. Whether or not these students constitute Type II errors (false negatives) need not concern us here. We simply note that this type of error has been traditionally deemed acceptable because all the student risks is additional instruction. Whatever the reason for the difference in classification from the different criteria, these cases are relatively rare. Even rarer are those students who master 15 objectives but score below 216. These are probably the students who consistently master only four of the

TABLE 3

Relative Frequency Distributions of Standardized Log Achievement Scores (SLAS) of Students Who Met and Who Did Not Meet MDE Criterion of Mastery of Fifteen Objectives on Grade Four MEAP Reading Test

SLAS	NON MASTERS	MASTERS
231 thur 258	-	40.7
230	-	9.0
229	-	-
228	-	7.8
227	-	5.8
226	-	-
225	-	5.2
224	-	5.5
223	-	5.5
222	0.4	4.3
221	0.4	4.3
220	0.4	4.1
219	1.1	2.9
218	2.4	1.8
217	5.3	2.2
216	2.3	0.3
215	3.0	0.3
214	5.4	0.1
213	2.8	0.1
212	4.9	-
211	5.1	-
210	5.0	-
162 thru 209	61.4	-
TOTAL PERCENT	99.9	99.9
MEAN	204.0	230.2
SD	10.2	9.0
MEDIAN	205.8	228.5
(N)	(998)	(1570)

TABLE 4

Relationship Between Mastery of 75 Percent of Grade Four MEAP Reading Objectives and Standardized Log Achievement Criterion Score Levels

(Percent of Total)

	Standardized Log Achievement Score		
	LE 215	GE 216	Total
Master 14 or fewer objectives	34.1 (875)	4.8 (123)	38.9 (998)
Master 15 or more objectives	.4 (9)	60.8 (1561)	61.1 (1570)
Total	34.1 (884)	65.6 (1684)	100.0 (2568)

five items for each objective and they constitute only about one half of one percent.[1] The point bi-serial correlation among these two criterion variables is .89. We feel safe in concluding that using SLAS criterion score of 216 enables us to make essentially the same mastery decisions as a mastery decision using 75 percent of the objectives.

What may be more informative than this summary discussion is the behavior of the SLAS distribution over the restricted range where mastery decisions are most difficult.[2] Table 3 indicated that mastery decisions are essentially being made in the score range 213 to 222. No student who mastered 15 or more objectives scored lower than 213 and no student who mastered at most 14 objectives scored higher than 222. It is in this region of "overlap" where precise measurement is most desirable. We note that the Rasch model is most efficient when the achievement level of the subjects are matched to the difficulty level of the items measuring their achievement. Less than 10 percent of the items from this test calibrate at the difficulty level which is near the region of "overlap" of these distributions. There are only eight items on the entire test with log item difficulties greater than 212 and only three of these items have difficulties greater than 216.[3]

---

[1]It is possible to master 15 objectives with a SLAS of 206, corresponding to a raw score of 60. In these data, the lowest SLAS achieved by students who mastered 15 objectives was 213.

[2]We believe that mastery decisions about students at the extremes of the distribution are relatively easier than those about students in the middle of the distribution. Table 3 indicated the "lumping" that occurs at the extremes. Over three-fifths of the non-masters (61.4 percent) fall in the first quartile of the total quartile of the total distribution of SLAS; two-fifths (40.7 percent) of the masters fall within the top quartile. Moreover, there are no masters in the lowest quartile of non-masters in the top quartile.

[3]Seven of the ten most difficult items on the test appear after test question number 88, suggesting that test order may be contributing to their difficulty level. We have not checked the rates of noncompletion for these items at this writing.

It is important to remember that the MEAP tests are designed to measure "minimal competencies." The fact that the competencies covered in the fourth grade reading test may be somewhat below what constitutes a typical fourth grader's battery of reading skills is indicated by the fact that the average achievement level of the students in our sample was 220 and the median was 222. These scores are considerably above the 200 average imposed by the calibration technique. What is troubling is the fact that so many students (65 percent score above 216) must take so many items that are so easy for them, resulting in scores that are of practically no instructional value, regardless of how they are reported.

In this section, we have demonstrated the essential similarity between the decisions about the mastery of students on the Grade Four MEAP Reading Test using the Rasch model - derived SLAS and 75 percent of the objectives mastered. For the majority of students, we found that the items were relatively easy given their achievement levels and that the amount of information available for instructional purposes was slight. In the following section, we explore an alternative to current testing practice which promises a significant reduction in the amount of testing without a loss in the information provided by current summary statistics.

The Rasch Model and Short Tests

The Rasch model offers a unique solution to the problem of statewide assessment of "minimal competencies." Under the assumptions of the Rasch model, measurement can be "test free." It is not necessary to administer all items to all students in order to make statements about whether the students have mastered certain "minimal competencies"--whether in terms of 19 (or 100) reading objectives or in terms of 95 (or 10,000) reading items. A student who receives a SLAS of 216 has met the criterion in terms of the content measured by Grade Four MEAP Reading. The power of the Rasch model lies in its ability to allow us to determine a student's SLAS by administering considerably fewer than 95 items. Once the items (or objectives) have been calibrated -- assigned a known difficulty level in relation to all the other items in the test -- all the items need not be administered to determine how students will perform on the skills that they measure.[1] The Rasch model allows the educator to measure skills without directly testing for them.

In order to determine empirically the ability of a short test to provide the same mastery information about students as longer tests, we developed a ten item test of fourth grade reading. The items were selected on the basis of the calibrations of items for the 95 item test. The items and their difficulty estimates are listed in Table 5.

---

[1]Brink (1972) demonstrated that since "the Rasch model scales items on easiness and subjects on achievement level," while "the Guttman model orders items on difficulty and the subjects on total score, " the Guttman model "does not possess the precision that may be possessed by a Rasch scale." We will not examine the underlying scability of the 95 items on the Grade Four MEAP Reading Test in this paper but will simply alert the reader to this property of the model.

TABLE 5 ABOUT HERE

The procedure used to identify items for the ten item test involved the identification of the 20 items with the highest difficulty estimates on the 95 item test. This list was then examined for those items with the best fit statistics, those primarily with FMS close to 1.00. Although the objectives with which the items were associated were not considered in their selection, we noted that seven objectives contributed items to the test, with one objective alone contributing three items: Objective No. 11 (see Appendix A). We noted also that five of these items are in the last 15 items that were administered, but MDE assures its user that the test is not speeded. Our choice of the twenty most difficult as the basis for the test rests on the consideration of the general level of easiness of the items relative to the subjects taking the test.

Having selected the items for the short test, we again set the criterion for mastery at 80 percent of the ten items and assigned a SLAS criterion score of 226 to the mastery decision.[1] We then arrayed the results of this sorting by percent mastery and SLAS on the 95 item test. We shall first consider the relationship between SLAS on the 95 item test and SLAS on the short test. Table 6 reports the results of the comparison of mastery according to a SLAS of 216 on the 95 item test and a SLAS of 226 on the ten item test. We see that there is a high correlation between the two criteria (r= .67). The mastery decisions agree in four fifths (81.0 percent) of the cases: about a third (33.8 percent) score below 216 on the 95 item test and below 226 on the 10 item test; almost one half (47.2 percent) score above the respective SLAS criterion scores on both tests.

[1] The SLAS criterion score is substantially higher for the short test since the average item difficulty is substantially higher. Techniques for equating the different length tests allow direct comparison of the performance of students on either form of the test (Rentz and Bashaw, 1975; Brigman and Bashaw, 1976).

TABLE 5

Item Statistics for Items Included in 10 Item Reading Test

Item Name	Item Diff[a]	Disc[b] Index	FMS[c]
177	.96	1.14	.97
189	1.22	1.22	.97
182	.79	1.10	.98
1100	.81	.95	1.03
199	1.44	.90	1.06
134	1.14	.81	1.07
197	1.36	.74	1.10
194	1.27	.73	1.11
162	1.31	.46	1.16
190	1.85	.58	1.18

[a]Rasch Log Item Difficulty estimates from 95 item calibration.

[b]Discrimination Index estimated from 95 item calibration.

[c]Fit Mean Square estimated from 95 item calibration.

What is particularly interesting is that although about one fifth (18.4 percent) of the students did not meet the criterion on the short test but

TABLE 6 ABOUT HERE

did meet the criterion of the 95 item test, less than one percent (0.6 percent) passed the short test and failed the longer version. The short test seems to impose a more rigid criterion than the longer test.

To make sense of the pattern in the off-diagonal cells in Table 6 we must consider the kinds of error that may be involved in making mastery decisions about students. Type I errors, false positives, involve deciding that a student has mastered the content tested when, in fact, he/she has not. Type II errors, false negatives, involve deciding that the student has not mastered the content when, infact, he/she has. If we assume that the results of the 95 item test are more believeable and accept that distribution as our picture of what is the case, the short test has caused 16 Type I errors and 473 Type II errors. If we, in addition, assume that Type I errors are more serious since the cost may include deciding not to provide additional instruction where it is needed, we find that the short test performed exceptionally well. Using one-tenth the amount of testing, there were almost no false positives. If the short test were used as a screening device for more exhaustive testing, the 473 Type II errors would be identified and corrected. Further, if the the purpose of additional testing was diagnostic, almost half the students could be exempted. The consequent reduction in interference with instruction and cost of administering tests would be considerable. At least in so far as the 95 item test represents a student's "true" level of reading skill, the short test would seem to perform adequately for making student mastery decisions.

TABLE 6

Relationship Between Standardized Log Achievement Criterion Scores on 95 Item and 10 Item Grade Four MEAP Reading Tests

(Percent of Total)

	Standardized Log Achievement Score 10 Item Test		
	LE 225	GE 226	Total
LE 215	33.8 (868)	0.6 (16)	34.4 (884)
GE 216	18.4 (473)	47.2 (1211)	65.6 (1684)
Total	52.2 (1341)	47.8 (1227)	100.0 (2568)

A similar result emerges when mastery decisions based upon the SLAS score on the ten item test are compared to those based upon mastery of 75 percent of the reading objectives. Table 7 shows agreement in 83.5 percent of the cases. Even fewer Type II errors (14.9 percent) appear and only

TABLE 7 ABOUT HERE

slightly more Type I errors (1.6 percent). Again, we find that the short test sorts students into mastery groups almost as efficiently as the longer versions of the test.

Conclusions

The present paper does not attempt to include any evaluation of either the MEAP Grade Four reading items, objectives, or reports. What we have attempted to show is that a relationship exists between number of objectives mastered, total test Standardized Log Achievement Score, and SLAS derived from a ten item subset of the 95 items. Our motivation for examining these relationships stems from three diverse areas of concern about the current praticies of MDE in the MEAP.

First, many districts find that there is little instructional use for MEAP results since nearly all of their students "master" nearly all of the objectives. These districts do, however, use MEAP results. They use them to show that their students are at least acquiring "minimal competencies." We are not in a position to evaluate this kind of use for the data. We simply believe that essentially the same information could be obtained by administering as few as five or ten items to students. Our analysis lends a great deal of support to this contention.

Second, with more and more local, state and federal programs requiring

TABLE 7

Relationship Between Standardized Log Achievement Criterion Score on 10 Item Test and Mastery of 75 Percent of Grade Four

(Percent of Total)

	Standardized Log Achievement Score 10 Item Test		
	LE 225	GE 226	Total
Master 14 or fewer objectives	37.3 (958)	1.6 (40)	38.9 (998)
Master 15 or more objectives	14.9 (383)	46.2 (1187)	61.1 (1570)
Total	52.2 (1341)	47.8 (1227)	100.0 (2568)

more and more evaluation data, testing time has become a major issue for many educators. It is very important that time which is devoted to testing be useful both for program evaluation as well as for instructional purposes. Under classical test theory, testing data for one purpose are usually not appropriate for the other. The Rasch model is a vehicle that provides a theoretical framework within which students may be tested with instruments appropriate for their achievement level, both in terms of content and difficulty, and yet which yields data for comparative analysis. Many responsible educators have proposed that some way be developed to allow local educational agencies flexibility in terms of the content and difficulty of the tests administered to their students. The current investigation suggests that a core of as few as ten test items from the present test could provide the MDE with essentially the same summary data on the attainment of minimal competencies as is currently available.

Third, if a statewide item bank (such as is being developed in MTSS) could be created following the Oregon model (which includes the Rasch item difficulty estimate for every item that is placed in the bank) the MDE could reduce the extent to which they might "dictate curriculum." Even within the context of testing for "minimal competencies," LEA's should be allowed to use achievement tests which reflect the content of their curriculum. When items of known difficulty which cover a broad range of content are made available to the educational community, LEA's will be able to test for what they teach and the MDE will be able to meaningfully summarize their data.

In summary, we believe that the approach outlined in this paper provides a way to enhance the utility of MEAP. If the implications of this investigation are acted upon, testing time could be drastically reduced while allowing for the testing of more diverse instructional

content. Further exploration of these techniques for application in test development and the establishment of criterion levels is needed. However, our findings here, and in other investigations in progress, suggest that the Rasch model is a very promising tool for understanding the results of criterion-referenced tests.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability, in F. Lord and M. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison Wesley, 1968; 397-479.

Brigman, S., Bashaw, W. L. Multiple test equating using the Rasch model. Paper presented at Annual Meetings of American Educational Research Association, San Francisco, 1976.

Brink, N. Rasch's logistic model vs. the Guttman Model. Journal of Educational Measurement, 1972, 32, 921-927.

Cypress, Beulah K. The effects of diverse test score distribution characteristics on the estimation of the ability parameter of the Rasch measurement model. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

Hambleton, R., Traub, R. Analysis of emprical data using two logistic latent trait models. Paper presented at Annual Meetings of American Educational Research Association, Minneapolis, 1970.

Kifer, E., Bramble, W. The Calibration of a criterion-referenced test. Paper presented at Annual Meeting of American Educational Research Association, Chicago, 1074.

Lord, F. M., Novick, M. R. Statistical Theories of Mental Test Scores. Reading, MA.: Addison-Wesley, 1968.

Lord, F. M. Quick estimates of the relative efficiency of true tests as a function of ability level. Journal of Educational Measurement, (Winter), 247-254.

Mead, R. Assessing the fit of data to the Rasch model. Paper presented at Annual Meeting of American Educatonal Research Association, San Francisco, 1976.

Michigan Educational Assessment Program. Technical Report, 1975-1976: Michigan Educational Assessment Program. Lansing, Michigan: Michigan Department of Education.

Do YOU Use MEAP Test Appropriately? Lansing, Michigan: Michigan Department of Education, no date.

Nie, N., et al. Statistical Package for the Social Sciences. Chicago: McGraw, 1975.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Denmark: Danmarks Paedogogiske Institute, 1960.

Rasch, G. An individualistic approach to item analysis, in Paul Lazarsfeld and Neil Henry, Readings in Mathematical Social Science. Chicago: Science Research Associates, 1966.

Rentz, R. R., Bashaw, W. L. The National reference scale for reading: An application of the Rasch Model. Jorunal of Educational Measurement, 14 (Summer), 161-180, 1977.

Rentz, R. R., Bashaw, W. L. Equating Reading Tests with the Rasch Model. Volume Q. Athens, GA: University of Georgia, Educational Research Lab., 1975.

Rudman, H. C. The use of data for decision-making. Michigan School Board Journal. (June), 12-13, 1977a.

Rudman, H. C. The Michigan Assessment Program, 1976-1977: The objectives. Michigan School Board Journal. (July), 10-11, 1977b.

Rudman, H. C. The Michigan Educational Assessment Program: The items.. Michigan School Board Journal. (August), 10-11, 1977c.

Rudman, H. C. The Michigan Educational Assessment Program, 1976-1977: Its meaning to school boards. Michigan School Board Journal. (September), 19-30, 1977d.

Tinsley, H., Dawis, R. An Investigation of the Rasch Simple Logistic Model: Sample-Free Stem and Test Calibration, Technical Report No. 3005. Washington, D.C.: Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, 1972.

Wright, B. Sample-free and person free measurement. Paper presented at the National Seminar on Adult Education Research, Chicago, 1968.

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

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

Wright, B., Mead, R. J. BICAL: Calibrating Items and Scales with the Rasch Model. Research Memorandum No. 23. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1977.

APPENDIX A

READING OBJECTIVES MEASURED IN THE 1977-78 MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM*

Grade 4

	Objective Number	
1.	2.1	Given a reading selection at the third grade level, the learner will match a series of words in the selection with appropriate definitions.
2.	2.2	Given a set of phrases, the student will indicate those phrases which have the same meaning.
3.	3.2	Given a reading selection at the third grade level in which every fifth word has been replaced with a blank, the learner will choose the exact word appropriate to the blank space at 50% accuracy.
4.	4.1	Given a method of arranging data, the learner will identify the method (e.g., color, size, importance, time, etc.)
5.	4.4	Given a series of randomly placed words, the learner will be able to alphabetize the words through the first three letters.
6.	5.1	Given a series of reading selections, the learner will indicate those which are factual.
7.	5.2	Given a series of reading selections, the learner will indicate those which are fictional.
8.	6.1-6.3	Given a reading selection. the learner will be able to identify the author's purpose (e.g., persuasion, entertainment, propaganda, etc.)
9.	7.1	Given a reading selection at the third grade level, the learner will select from a list of possible titles the one most appropriate as the title for that selection.
10.	7.2	Given a reading selection at the third grade level, the learner will select from a series of still pictures the one picture most appropriate in depicting the main idea of the selection.
11.	7.3	Given a reading selection at the third grade level, the learner will select from a number of short summaries the one which best summarizes the selection.

*This list contains only the objectives which are included in the every-pupil portion of the 1977-78 MEAP tests. A complete set of the objectives is available in **Minimal Performance Objectives for Communication Skills Education in Michigan,** Michigan Department of Education.

12.	8.4	Given a reading selection at the third grade level, the learner will match a series of direct quotations from the story with the character who is speaking.
13.	10.3	Given a reading selection at the third grade level, the learner will choose from a series of sentences that sentence which best describes how a given character feels in a story.
14.	10.6	Given a selection containing figurative language, the learner will identify from a series of descriptive phrases the phrase that most accurately describes the mood expressed in the selection.
15.	11.1	Given a reading selection at the third grade level, the learner will correctly match a series of causes with a corresponding series of effects.
16.	11.2	Given a reading selection at the third grade level with the conclusion of the story deleted, the learner will select from a series of possible conclusions the one most appropriate to the selection.
17.	13.1	Given a locational question, the learner will choose from a series of reference sources where that item will be found.
18.	13.2	Given a locational question about newspapers, the learner will select the section where the answer would be found.
19.	14.1-14.3	Given a reading selection at the third grade level, the learner will answer correctly a series of multiple choice questions relating to meanings, generalizations, or conclusions not expressed in the selection itself.

LIST OF ITEMS MEASURING EACH FOURTH GRADE OBJECTIVE

Reading

Objective Number	Item Number
1	45,52,78,81,92
2	83-87
3	65-69
4	16-20
5	6-10
6	27-31
7	35-39
8	24,32,33,76,98
9	41,53,74,89,97
10	21,40,51,70.96
11	34,43,80,90,99
12	42,48,72,77,88
13	47,49,75,79,93
14	11-15
15	23,44,50,91,100
16	22,46,71,82,95
17	55-59

Mathematics

Objective Number	Item Number
1	106-200
2	101-105
3	241-245
4	231-235
5	226-230
6	136-140
7	176-180
8	246-250
9	111-115
10	166-170
11	116-120
12	156-160
13	151-155
14	146-150
15	236-240
16	191-195
17	121-125
18	171-175
19	211-215
20	251-255
21	106-110
22	161-165
23	1-5
24	206-210
25	126-130
26	201-205
27	141-145
28	186-190
29	216-220
30	221-225
31	256-260
32	181-185
33	131-135