R E P O R T R E S U M E S

ED 012 211 RC 001 478

ANALYSIS OF THE TESTING PROGRAM - MDTA PROJECT.

BY- PEARCE, FRANK C.

PUB DATE 65

EDRS PRICE MF-$0.09 HC-$0.72 18P.

DESCRIPTORS- *ACHIEVEMENT TESTS, CURRICULUM, CURRICULUM DEVELOPMENT, *PROGRAM EVALUATION, *OCCUPATIONS, *TESTING PROGRAMS, TESTING, *TEST INTERPRETATION, TEST CONSTRUCTION, STANISLAUS COUNTY MULTI OCCUPATIONAL PROJECT, CALIFORNIA ACHIEVEMENT TEST, MODESTO

A DISCUSSION OF TEST VARIABLES AND THE TESTING PROGRAM OF THE PROJECT WAS INITIALLY PRESENTED IN THIS DOCUMENT. CAUTION WAS PARTICULARLY EMPHASIZED IN THE INTERPRETATION OF RESULTS USING STANDARD ACHIEVEMENT TESTS WITH SUBJECTS FROM THE MODESTO MULTI OCCUPATIONAL PROJECT. RECOMMENDATIONS FOR IMPROVED PROGRAM EVALUATION INCLUDED ADDITIONAL CURRICULUM REVISION, IMPROVED TESTING PROCEDURES, AND CHANGES IN TEST CONSTRUCTION. (JS)

Modesto Multi Occupational Project

RESEARCH REPORT NO. 3

Subject: Analysis of the Testing Program - MDTA Project

Date: December 28, 1965

[Frank C. Pearce]

During the past several months considerable concern has been expressed by members of the staff about trends they felt existed in the testing program. It is the purpose of this report to explore the validity of some of the generalizations which have been made. This report deals specifically with the California Achievement Test (CAT) and includes a general discussion of testing, empirical test data, interpretations and recommendations.

A Testing Program

Achievement tests are used to measure an individual's present level of knowledge, skills, and competence. They do not measure intelligence, predict a student's future performance (aptitude), or indicate areas of interest. They only measure how well a student can perform some task such as reading, spelling, or solving math problems. How well any given student will do depends on his innate ability, his acquired ability, and his motivation.

Innate ability is usually equated with the term intelligence, and, as such, is not considered to be a characteristic which the instructor can manipulate. Thus, if a trainee does not possess the necessary mental ability he cannot be expected to do well or generally change his test scores in any way. Any change which does occur in his scores will probably be the result of chance alone.

Acquired ability is the skill, or learning, that the teacher has helped the student to attain.

Motivation here is the desire to do well on a test. It would seem appropriate to suggest that some trainees exhibit very little, if any, desire to do well on the test. In fact, they may have performed poorly intentionally.

Probably the single most important criterion for judging a test is its validity. How well is it able to measure what it is supposed to measure? There are several kinds of validity, but only two are discussed here: (1) face validity, and (2) content validity.

Face validity simply means whether the test looks, to both the teacher and the trainee, as if it measures what is being taught in the classroom. It is doubtful that the Modesto trainee is being taught the meaning of the word servitude, how to make an adjective from a noun, how to find the area of a parallelogram, or how to spell melancholy--all of which are a part of the test. Even if an attempt were being made to teach these things, it is questionable that the trainee would consider them important enough to remember. Thus, the face validity of the test is problematic. It is generally recognized that face validity contributes to motivation, since people try harder when the test seems reasonable. This, then, is a facet which would seem to require some improvement.

Content validity indicates how well the test covers the important points of a training program. This kind of validity is particularly important when it is recognized that content for the test was selected from school curriculums across the country. What relationship, if any, is there between the Modesto program's curriculum and that of the traditional public school?

Next, consider the matter of norms, or normal performance on the test. The norms for this test are excellent, perhaps some of the best in the area of standardized tests, but they are based on elementary, junior high and high school student performance. The purpose of a norm is to establish a basis upon which to compare the scores of a person taking the test with the scores of others within the group of which he is a part. Can the score of an adult from a deprived background be compared with the norm population provided by the test authors? Should norms be established specifically for the project to reflect the characteristics of the trainees?

Other factors influencing the test scores include (1) time limits, and (2) the method of recording answers. Working under time limits increases the trainee's anxiety to the point where it interfers with test performance, especially since he is apprehensive at best and far from being test wise. In fact, some trainees feel that if they do not do well on the test they let the Modesto project down; others are afraid that they may not be able to enter some vocation if they are not successful. The method of recording answers to the questions introduces the possibility of clerical errors for those who are not test wise. This, too, can cause wide variations in the test scores.

The above comments are not directed at the test authors, since the validity of the CAT, in terms of the purposes for which it was designed, is high. However, caution is urged for those who wish to interpret CAT scores in terms of MDTA trainees. Recognition must be given to the test limitations and how they influence test results. A test is a tool that can be of great help if used properly, or of considerable harm when used incorrectly.

The following data include a Grade Placement score (GP), which can be misleading. The Grade Placement score is given in tenths of a school year, i.e., 5.6 means the sixth month of grade five, 7.1 the first month of grade seven. The Grade Placement score assumes that subjects are taught uniformly throughout the school year, which probably is not done, particularly in the Modesto project.

Given any particular trainee who enters the program at some definite point in time, it is possible that a specific concept measured by the test was taught prior to entry and will not be taught again until after the trainee has left the project. Thus, a portion of the trainee's Grade Placement score is missing, no matter what he does. Moreover, Grade Placement scores are based on averages and, therefore, require knowledge of the standard deviation or standard error of measurement that is involved in a test score. Neither of these are available.

Next, the period of time during which a trainee can be exposed to pre-vocational training is somewhat limited, which means that the breadth and depth of trainee understanding cannot be adequately related to a Grade Placement score. In view of these and other considerations, it may be appropriate to use some standard score as the Z score.

What the Test Data Shows

For those who are interested in the test scores for each person for each test, differences from one test to the next, and the time between tests, the information is available. This section summarizes that data.

Reading

Table I shows the number of persons whose test scores in reading have increased or decreased between their first and second test and their second and third test. There were no apparent differences between the number of persons whose vocabulary scores increased or decreased. A significant number of persons increased their reading and comprehension scores (49% more increased than decreased) from one test to the next, while 41% more trainees increased total reading scores than decreased. These results suggested that there were no real changes in vocabulary skills, but there was an increase in reading comprehension and total reading scores between tests one and two which was considerably above that which could be expected by chance alone. There were no differences in any of the reading areas between tests two and three.

TABLE I CHANGES IN READING PLACEMENT SCORES

Test I vs Test II

Change	Vocabulary		Comprehension		Total	
	f	%	f	%	f	%
Increase	44	50	62	71	58	67
No Change	6	7	6	7	6	7
Decrease	37	43	19	22	23	26
Total	87	100	87	100	87	100

Test II vs Test III

Change	Vocabulary		Comprehension		Total	
	f	%	f	%	f	%
Increase	9	41	14	64	8	36
No Change	1	5	0	0	3	14
Decrease	12	54	8	36	11	50
Total	22	100	22	100	22	100

When the mean change in Grade Placement scores for each trainee is examined the trends are not quite as clear. There was an increase in vocabulary of 0.1 of a Grade Placement score, an increase of 0.4 (approximately one-half year) in reading comprehension, and an increase of 0.3 of a Grade Placement score in total reading between test one and test two. This indicated that the trend toward an increase in the area of reading from one test to the next was more apparent than real, although changes were in a positive direction. There were no real differences in mean Grade Placement scores between the second and third test. These changes in reading scores occured within a time period of one to eleven months, with a mean of four months.

Mathematics

Table II indicates the number of persons whose scores increased or decreased from one test to the next. A significant difference in terms of the number of persons whose Grade Placement scores increased over those who decreased between the first and second test was found in all areas of mathematics. These results were repeated between the second and third tests, with the exception of differences in mechanics of mathematics. Differences in mechanics of mathematics were no more than would occur by chance alone.

TABLE II CHANGES IN MATHEMATICS PLACEMENT SCORES

Test I vs Test II

Change	Reasoning		Fundamentals		Total	
	f	%	f	%	f	%
Increase	55	76	60	83	60	83
No Change	1	2	1	2	2	3
Decrease	16	22	11	15	10	14
Total	72	100	72	100	72	100

TABLE II (CONTINUED)

Test II vs Test III

Change	Reasoning		Fundamentals		Total	
	f	%	f	%	f	%
Increase	18	67	14	52	18	67
No Change	0	0	2	7	2	7
Decrease	9	33	11	41	7	26
Total	27	100	27	100	27	100

In terms of changes in mean Grade Placement scores, no differences were found in reasoning, fundamentals, or toal mathematics scores between the second and third tests. However, between the first and second test a specific increase for each trainee was noted. A mean increase of 0.5 was found in reasoning, 0.6 in fundamentals, and an 0.6 mean increase in Grade Placement score in total mathematics. The average elapsed time was four months, with a range from one to eleven months. Thus, it would appear that an increase of one-half year Grade Placement score could be expected in mathematics every four months.

Language

Table III indicates the increases or decreases in language Grade Placement scores for Modesto trainees.

TABLE III CHANGES IN LANGUAGE PLACEMENT SCORES

Test I vs Test II

Change	Mechanics		Spelling		Total	
	f	%	f	%	f	%
Increase	47	64	45	61	47	64
No Change	4	5	4	5	3	4
Decrease	23	31	25	34	24	32
Total	74	100	74	100	74	100

TABLE III (CONTINUED)

Test II vs Test III

Change	Mechanics		Spelling		Total	
	f	%	f	%	f	%
Increase	11	73	8	53	11	73
No Change	0	0	0	0	0	0
Decrease	4	27	7	47	4	27
Total	15	100	15	100	15	100

There was a significant difference in the number of persons whose language scores increased over those whose scores decreased in all language areas between the first and second test. Increase in trainee's scores in mechanics were 33% more than decrease, 27% more in spelling, and 32% more in total language. Similar results were found between tests one and two with the exception of spelling, where no real difference was found. This trend was also found for the mean Grade Placement scores between tests one and two, but the difference in total language score was the only one which was significant.

The analysis of test data indicated that very few people remained at the same Grade Placement score level from one test to the next. That is, there were few people whose test scores did not change. However, Table V indicated that one-third of those taking the CAT changed 0.4 or less of a Grade Placement score between the first and second test. This change could have been an increase or decrease, but the total change was less than 0.4 of a school year. Moreover, similar findings were recorded between the second and third test. This would seem to suggest that 30% of the trainees exhibited very

little change from one test to the next in the skill areas measured. Conversely, 66% of the trainees were able to change their scores by one-half a school year or more. The important consideration here is how many were able to change in a positive direction.

TABLE V NUMBER OF PERSONS WITH GRADE PLACEMENT SCORES OF 0.0 to 0.4*

Test	Test I vs Test II		Test II vs Test III	
	f	%	f	%
Vocabulary	25	29	8	36
Comprehension	25	29	7	32
Total Reading	37	42	11	50
Reasoning	24	33	9	33
Fundamentals	16	22	19	52
Total Mathematics	20	28	13	48
Mechanics	19	26	4	27
Spelling	21	28	1	7
Total Language	25	34	4	27
TOTAL BATTERY	22	38	5	50

*Note that Tables I - IV provide the totals from which these percentages were derived.

By consulting Table VI it was found that generally 50% of the trainees increased their Grade Placement score by one-half of a school year between test periods. Specifically, 66% of the trainees demonstrated an increase of more than a 0.5 Grade Placement score in all areas of mathematics, while 60% of the trainees increased their total battery scores by one-half a year or more.

TABLE VI NUMBER OF PERSONS WITH GRADE PLACEMENT SCORES AT OR ABOVE THE 0.5 LEVEL BETWEEN TEST I AND TEST II

Test	0.5 to 1.6 Level		1.7 Level or Above		Total	
	f	%	f	%	f	%
Vocabulary	27	31	9	10	36	41
Comprehension	36	41	14	16	50	57
Total Reading	31	36	7	8	38	44
Reasoning	28	39	11	15	39	54
Fundamentals	31	43	21	29	52	72
Total Mathematics	38	53	11	15	49	68
Mechanics	31	42	9	12	40	54
Spelling	28	38	7	9	35	47
Total language	33	31	11	15	34	46
TOTAL BATTERY	29	50	6	10	35	60

The data from Tables V and VI suggested the following composite picture of the trainee population: 30% failed to demonstrate significant score changes; 20% demonstrated a decrease in Grade Placement scores of one-half year or more; 50% increased scores by one-half year or more, and all of these changes occurred within an average period of four months. It was found that of the 20% whose scores decreased by more than one-half year only 22 persons demonstrated a decrease as much as a 1.5 grade placement score. In addition, all but six of these trainees did so only on one of the sub-tests. This latter point is particularly important since the other scores are within the reported standard deviation.

Finally, one of the more important findings revealed through the test analysis was the amount of time required by the trainees to increase the level of their basic skill development. That is, the mean increase was 0.8, or approximately one school year in four months of basic education training. This compares with a report by Imel in 1965 on the San Diego basic education program where an increase of 2.0 grades occurred after 100 hours.

It was not possible to determine how many school days this 100 hours represented. In addition, Imel carefully points out the limitations of "attempting to cram several years of education into a few weeks" and it would seem unlikely that the 100 hours represented less than three months. The tests used to measure change were the Stanford and California Achievement Tests (intermediate forms) which tend to contribute to the comparability. The description of this program's curriculum can be generally compared with Modesto, although it was not clear whether motivation and attitudes are as integral a part of the program as they are at Modesto. Moreover, the goals of the two programs are not the same, since Modesto's curriculum is designed to prepare a person for vocational training. Thus, the comparability of results is somewhat tenuous; however, they are the best currently available.

Information on other efforts is limited, but several are included here for your information. Levi (1964) reports on a Chicago program that found an increase of 4.6 (S.D. = .65) after 99 hours of basic education. This figure is not comparable, however, since the meaning of the score is not clear. Moreover, it was not possible to determine

the similarities, if any, between the trainees and the curriculums. Specific techniques have also been reported to produce improvements of about two grades. These include a 1964 report by Henny using the "Family Phonics System" and a 1961 report by Allen using the "Laubach Literacy Films." The educational, psychological, and sociological journals report on other techniques such as the use of television, other parts of the Laubach series, the use of groups, etc., but in all cases (except the San Diego Report) it is practically impossible to compare results with those of the Modesto project.

Conclusions and Recommendations

1. Through an analysis of the data available it was evident that a hypothesis which stated that trainees were decreasing in skill development in reading and language could not be generally supported. Clearly, the majority of the trainees did improve the achievement scores they attained in reading, mathematics, and language.
2. The findings indicated that trainees needed additional emphasis in the classroom on vocabulary and reading skills as measured by the California Achievement Test. The empirical data indicated a continual improvement in all other test areas.
3. One of the more suggestive findings was concerned with the time needed to demonstrate a change in the trainee's skill development in reading, mathematics, and language. It is recommended that the Modesto project explore this entire topic and its ramifications in considerable detail. Although the increases noted in a given time period are meaningful, improvements can be made.

It would be well to explore:

a. The area of planning for classroom presentation.

b. The basic goals of each curricular area.

c. The emphasis to be placed on certain areas of the curriculum.

d. The proportion of time to be devoted to the 3 R areas and motivation-attitudes.

e. What areas should be measured as progress toward vocational goals.

f. Teaching the whole student vs. preparation for specific vocational areas.

g. The use of specific teaching techniques such as large groups vs. small groups, programmed learning, other audio-visual aids and varying uses of those available, team teaching, etc.

h. The concepts to be taught in each curricular area.

i. The manner in which communication, integration, and cooperation within the various curricular areas may be enhanced.

j. The manner in which flexibility, opportunities without administrative influence, the evolvement of new ideas, procedures and general innovation possible can be used to greater advantage. In fact, could the instructors conduct their own research in these areas?

4. There were not sufficient numbers of persons who had been tested a third time to establish whether or not the trend found between tests one and two were actually continuing.

5. Very little growth, if any, was noted for 33% of the trainees. Thus, it would seem appropriate to examine carefully the

individuals involved and suggest program or individual modifications which would alter this condition. This same point should be considered for the few whose scores decreased significantly.

6. Relatively few fluctuations in scores occurred which could not have been predicted. However, some exceeded the range of a single standard deviation. Connecting this fact with the knowledge that certain errors in the administration of the test materialized, suggests some modifications in the testing procedure. It is recommended that test administration become the responsibility of one or two persons and that the size of groups tested be limited to twenty or twenty-five persons per administrator. It is also recommended that all test scoring be handled by the 1230 interpreter and the results returned by the research section to the appropriate person (s).

7. A number of persons took tests at a level which was too high to provide an achievement measure. The procedure used to select test levels for specific trainees should be reviewed.

8. Discussion introducing this report clearly indicates the need for seriously questioning the use of the California Achievement Test. On the other hand, this criticism would apply to nearly any other standardized test. If the criticisms directed at the test's validity are in themselves valid, the program would have to develop its own instrument for measuring growth. If this alternative were selected it must be recognized that it contains a host of very difficult problems. An alternative to

this would be to develop norms for the Modesto project and to use a standard score such as the Z score. The latter alternative is recommended as the more profitable in terms of time and available abilities.

9. The anxiety and threat created within student trainees of the testing situation may well produce scores which do not accurately reflect the trainee's ability. One possible way to correct, at least partially, this condition would be to eliminate the time element. This action would violate the principle of maximum performance, but it could be, at least partially, negated through the development of local norms that eliminated the time element.

10. A definite procedure for involving the individual instructors in the use of test results is recommended. Comments by instructors clearly indicate a lack of information and understanding in this area. It would seem appropriate to consider the development of and in-service program around this specific test, or at least this test area.

11. It is recommended that individuals be tested three weeks after entering the program and every three months thereafter. It is also recommended that the research section provide the test administrators with the following: (1) date the test should be given, (2) the specific sub-tests or battery to be used, and (3) the form of test to be given. The level of test to be administered should be determined by the counselor. This would tend to avoid errors of repetition and omission.

12. In view of the misinformation and anxiety indicated by the remarks made by some students, the procedure for informing students

of the purpose of the test should be examined for possible means of improvement. Moreover, the results of this analysis raise the possibility that some trainees are purposefully doing poorly on the test. At the very least it would appear that the face validity of the test tends to inhibit their motivation to perform at full capacity.

If the trainee does not see the relationship between his vocational-perceived needs and the questions asked in the test, his level of motivation may be quite low or perhaps non existent. Additional explanation to the trainees of the purpose of testing is recommended.

13. Clearly a number of the scores earned by trainees are inaccurate because of trainee clerical errors during the test period. Failure to mark the proper bar for the corresponding question causes considerable loss in time if and when the error is discovered. It is likely that questions are marked as being incorrect when actually many may have been answered correctly. In any case the student who is inexperienced in the taking of tests is unfairly penalized. A change in the method used to record a response to a question should be considered or practice in the necessary technique should be given.

14. Finally, it would seem that this report clearly indicates the danger of generalizing beyond the available facts It is seldom indeed, when one can consider any variable all black or all white when that variable is dependent upon human behavior. This writer maintains that conclusions based on general impressions or imperical data are of value only when they provide alternatives or direction to modify and generally improve the program.