

a

BUROS ASSIGNMENT

A PROJECT

Presented to

Dr. Hyman

This project is presented  
in partial fulfillment  
of the requirements for the course  
Tests in Counseling and Guidance

by

Asa G. Hilliard

October, 1962



## INTRODUCTION

It was the purpose of this project to select a test for the measurement of junior high school achievement in school subjects at the end of grade nine. The test is to be used for academic counseling.

It may be that the appearance of the problem is more contrived than it should be for this assignment. This is probably due to the fact that I wanted especially to examine the reviews of the three tests under consideration. However, I am operating under the assumption that the main purpose of the assignment was to give us practice in the use of Buros.

Also, it may be noted that two of the three tests considered were from the Fourth Mental Measurements Yearbook rather than from the fifth. This happened because the fifth yearbook was in use on two of the occasions when it was needed. Nevertheless, I do understand that the most recent publication should be consulted.

In a real situation, it would be well to read each yearbook in which a test is reviewed. This would tend to show the attention given by publishers to the reviews. I found one case in which the reviewers were highly critical of a test publisher for failing to heed the suggestions of an earlier reviewer.

After a preliminary perusal of the section in Buros dealing with achievement tests, the following tests were selected as having possibilities; The California Achievement Tests, The Iowa Tests of Educational Development, and The College Entrance Examination Board's test in history.

## THE CALIFORNIA ACHIEVEMENT TESTS

The California Achievement Tests were intended as a revision of the Progressive Achievement Tests. It covers the school years from 1 to 14. The first test was published in 1933 and the last revision was in 1951. Ten scores are given. They are reading, vocabulary, reading comprehension, total reading, arithmetic fundamentals, and a total score. A condensed score can be given for reading, arithmetic and language. The tests can be scored by hand or by IBM. There are four forms of the test since 1950. Administration time is 165 minutes. The tests were reviewed by Warren Findley, Director of Test Development ETS; Alvin Schindler, Professor of Education, University of Maryland; and J. Harland Shorse, Professor of Education, University of Illinois. The treatment of the is for the INTERMEDIATE (grades 7-9) battery only.

Grade scores and percentile norms for the test were based on a sample of from 50,000 to 100,000. These are given for each grade level. The sample is stratified with respect to I.Q. distribution, amount of acceleration or retardation, proportion of caucasian, Mexican, Negro and other ethnic groups. Form AA was standardized on the eight geographical districts of the United States commonly used by test makers. The other three forms were made equivalent to form AA, i.e. items of each form had corresponding content and statistically determined difficulty. The other forms were assumed to be appropriately interpreted by the same norms after this procedure.

Findley found that the range of objectives measured was broad.



Shorse and Schindler felt that some important objectives had not been measured. Presumably these objectives would be of the "critical thinking and interpretation" type. Findley questioned the paucity of items for each objective. Schindler found that the ceiling was not high enough for junior high. Findley found that the material was suitable for the grade levels. However, he questioned the interpretation of reading, language, and arithmetic scores when these are not always taught in junior high. When the test items were built in 1933, they were based upon an analysis of courses of study in "representative" cities. Shorse is critical of the failure of published materials to indicate if this procedure for item selection has been continued. Shorse also notes that parts of the reading comprehension test can be answered without reading the passage. All three reviewers were highly critical of the continued claim by the publisher that the test is a "diagnostic" instrument. The basis for their objection is that the few items to measure each objective are insufficient and have no known reliability. The use of the word "analysis" in place of diagnosis would be more realistic. One further note on validity is made by Findley. The 90 items in vocabulary are divided into math, science, social science, and reading. However, these areas are highly correlated for vocabulary and the sub-scores are unreliable.

The reliability for a single grade is reported as being from .83 to .96 for the components of the test. Findley points out that the rule allowing 90% to finish the test may have affected the reliability of the tests because of the method of sampling. For example, a bright pupil in a dull class will have much more time than he would have had in a bright class where 90% finish



much faster. Shorse felt that it was unfortunate that the .83 coefficient was on the most important test of all, reading comprehension.

The intermediate test is thought by Findley to be the best of the batteries. The format is judged to be clear and good. All of the reviewers praised the ease of administration and scoring. The scoreeze answer sheet looks identical to the IBM answer sheet. The fact that all scoring is objective makes the test easier to score. A profile sheet is given and may be useful subject to the limitations discussed above under diagnosis. The test costs only \$3 per 25. Answer sheets for IBM are 4 cents each. For the scoreeze the price is 7 cents per sheet. The publisher is reputable. Buros lists very few research studies on this instrument.

#### THE IOWA TESTS OF EDUCATIONAL DEVELOPMENT

The Iowa tests of Educational Development are published by Science Research Associates. They are for use with grades 8.5 to 13.5. Publication started in 1942 and the last revision was in 1948. Ten scores are given on 9 subtests. The subtests are, basic social concepts, general background in the natural sciences, ability to interpret reading in social studies, ability to interpret literary materials, general vocabulary, and uses of sources of information. The ninth score is a total of the first seven. There are two forms of the test. A separate answer sheet must be used. The administrator must agree to use all of the pupils in a high school or in a grade level. Scoring, tabulation, and reporting of results is included in the rental

price, which is not reported in Buros. Administration time is 480 minutes. The test is reviewed by Eric Gardner, Associate Professor of Education, Syracuse University.

The percentile norms are based on 50,000 pupils in 290 high schools in Iowa and neighboring states. The use of such a geographically restricted sample was justified by the following facts. The median performance and inter quartile range on the Iowa Tests of Educational Development of a group of 30,000 graduating seniors selected by the Armed Forces as a national stratified sample corresponds closely to the Iowa portion of the sample. The inter-test correlation goes from .397 to .791.

The reliability of the subtests for pupils in a single grade in a single school goes from .81 to .94. This is a split-half  $r$  corrected by the Spearman-Brown Profecy Formula. Gardner says the forms "appear" to be similar but there is no evidence of a statistical test of this fact.

The test is not based upon the analysis of specific courses of study. It is based largely upon the things an educated person "needs" to know. This is determined by the test constructor. No statement of his frame of reference is made in the manual. Gardner thinks this must be done before results can be adequately interpreted. Face validity was the major basis for item selection, although item discrimination was also used. The new manual gives validity figures for the test compared to success in the first year of college.

The test format is judged to be good. The directions are well done. Item construction is good. The unique method of scaling used drew the praise of Gardner. The probable error interval was used for scale units in order to



help the teacher be cautious in interpretation. However, the fact that the probable error may change from level to level and from one part of the scale to another was pointed out by Gardner.

The publisher is reputable. There are very few studies reported in Buros that deal with the Iowa Tests of Educational Development.

#### THE CEEB SOCIAL SCIENCE TEST

The College Board tests are published by the educational testing service. Grades 10 and 11 can take the tests. The tests are scored by ETS. One test can be taken alone. IBM answersheets are used. Publication started in 1937 and continued through the last revision in 1958. The cost of the achievement tests is \$9. One afternoon is required for administration. However, 80 minutes is the length of the social studies test alone. Ralph Tyler, University of Chicago, reviews the test of social studies.

The test deals with United States History and not with social studies, as seen by Tyler. Of 100 items, 20 are based upon major movements in history such as the Protestant Revolt, TVA, and dollar diplomacy. Twenty items are based upon people connected with certain policies. The remaining items sample knowledge of events. Tyler is critical of the failure of the test to require knowledge of dependable sources of information on social phenomena. Further, history is given as fact and the problems of interpretation are glossed over.

Tyler states that the test has a reliability of .92. He does not tell



how the figure was obtained.

The topography is judged to be good. The test is well edited. Administration and scoring are easily accomplished.

#### SUMMARY

As far as the mechanics of test construction are concerned, the CAT seems to be most desirable. However the objectives tested by the Iowa seem to be closer to general educational philosophy. This test seems to lack the needed work on mechanics. The CEEB needs to be reworked in order to measure a broader range of social studies objectives. It was hoped that it would correct this deficiency in the other two tests and, therefore, could be used in conjunction with one of them. However, history is not social studies. If I had to choose between the three tests above, I would select the CAT.