DOCUMENT RESUME

ED 060 135 TM 001 401

AUTHOR Ebel, Robert L.
TITLE 1971 AERA Conference Summaries: IV. Test Development, Interpretation, and Use.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
REPORT NO TM-R-14
PUB DATE Mar 72
NOTE 17p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS College Students; *Creativity Tests; Disadvantaged Youth; Educational Testing; Elementary Education; Factor Analysis; Intelligence Tests; Multiple Choice Tests; *Performance Factors; Predictive Validity; Response Style (Tests); Secondary Education; Student Characteristics; *Test Construction; *Testing; Testing Programs; Test Interpretation; *Test Validity
IDENTIFIERS AERA; *American Educational Research Association

ABSTRACT

This report presents a thematic summary of the AERA papers dealing with test development, interpretation, and use, presented at the 1971 meeting in New York City. Papers were grouped into the following categories: test development and validation; inventory development and validation; measurement of creativity; factors in test performance; use of tests to measure status or change; use of tests to predict; and use of tests to foster learning. (AG)

TM REPORTS

NUMBER 14

1971 AERA Conference Summaries

IV. Test Development, Interpretation, and Use

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION ■ EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540
Conducted by Educational Testing Service in Association with Rutgers University Graduate School of Education

The Clearinghouse operates under contract with the U. S. Department of Health, Education and Welfare, Office of Education. Contractors are encouraged to express freely their judgment in professional and technical matters. Points of view expressed within do not necessarily, therefore, represent the opinions or policy of any agency of the United States Government.

March 1972

TEST DEVELOPMENT, INTERPRETATION, AND USE

Robert L. Ebel

ERIC Clearinghouse on Tests, Measurement, and Evaluation

PREVIOUS TITLES IN THIS SERIES

1. Developing Criterion-Referenced Tests
   ED 041 052

2. Test Bias: A Bibliography
   ED 051 312

3. Ability Grouping: Status, Impact, and Alternatives
   ED 052 260

4. Developing Performance Tests for Classroom Evaluation
   ED 052 259

5. Tests of Basic Learning for Adults: An Annotated Bibliography
   TM 000 987 (ED number not yet available)

6. State Educational Assessment Programs: An Overview
   TM 001 024 (ED number not yet available)

7. Criterion Referenced Measurement: A Bibliography
   TM 001 046 (ED number not yet available)

INTRODUCTION

About 575 of the 700 papers presented at the 1971 AERA Annual Meeting in New York City were collected by the ERIC Clearinghouse on Tests, Measurement, and Evaluation (ERIC/TM). ERIC/TM indexed and abstracted for announcement in Research in Education (RIE) 175 papers which fell within our area of interest - testing, measurement, and evaluation. The remaining papers were distributed to the other Clearinghouses in the ERIC system for processing.

Because of an interest in thematic summaries of AERA papers on the part of a large segment of ERIC/TM users, we decided to invite a group of authors to assist us in producing such a series based on the materials processed for RIE by our Clearinghouse. Five topics were chosen for the series: Criterion Referenced Measurement, Evaluation, Innovation in Measurement, Statistics, and Test Construction.

Individual papers referred to in this summary may be obtained in either hard copy or microfiche form from:

ERIC Document Reproduction Service (EDRS)
P. O. Drawer O
Bethesda, Maryland 20014

Prices and ordering information for these documents may be found in any current issue of Research in Education.

Editor, ERIC/TM

This summary attempts to organize 39 papers on tests, measurement, and evaluation into related categories, and to briefly summarize each. The following categories were adopted: Test development and validation (7 papers); Inventory development and validation (7 papers); Measurement of creativity (5 papers); Factors in test performance (7 papers); Use of tests to measure status or change (6 papers); Use of tests to predict (4 papers); and Use of tests to foster learning (3 papers).

Occasional evaluative reactions are implied in the brief summaries of each paper, and some general comments on the quality of the papers are made in the concluding section.

A. Test Development and Validation

The two studies in this group which might be called test development studies did not involve the development of new tests. Rather, they were concerned, in one case, with adapting an existing test for computer administration and scoring, and in the other, with improving the reliability of an existing test.

Heal developed and evaluated a program for administering and scoring the *Slosson Intelligence Test*, using the IBM 1500 Instructional System. Correlations in the order of r=.75 (somewhat lower than desired) between automated and conventional administrations of the test, were explained in part by the homegeniety of the sample of 48 college students used in the study. The investigators see promise that the procedure may increase the predictive efficiency of the test by making it less threatening to examinees, and by providing data on response latency and other indices of ability.

An attempt to improve the reliability of the *Sigel Cognitive Style Test* was reported by Scott. The test consists of thirty-five cards, each bearing three pictures of familiar objects. The examinee's task is to select a pair of pictures and justify his selection. After item analysis two shortened forms, one for males, the other for females, were developed. These forms were shown to have somewhat higher reliability than the original test, though the need to use two forms tends to complicate test administration.

A variety of methods for test validation were illustrated in the five studies of this type in the present group. Fischbach undertook to check

the validity of a test of word attack skills, one presumed component of general reading skill. His study showed the expected generally positive, monotonic relation between scores on the general reading test and number of word attack skills mastered.

To investigate the validity of true-false test items as alternatives to multiple-choice items, Ebel converted multiple-choice science test items to true-false form, administered both forms to a group of college students, and correlated the resulting scores. When corrected for attenuation, the correlations indicated that the two forms can provide nearly identical measures of achievement. A larger study would have provided more dependable evidence.

Do teachers who score high on a test of competence in judging the quality of student writing agree with each other in rating themes? Whalen found that they agreed moderately well (r=.79). But those who scored low on the test of competence agreed with each other and with the high scores even better (r=.84). This casts some doubt on the validity of the test of competence.

Bierly investigated the validity of a test which required young children to manipulate concrete materials in response to an adult experimenter's utterances. She found the test to yield highly reliable scores, and to differentiate Headstart and Day Care Center children. Sentences in the active voice yielded more correct responses than sentences in the passive voice. No other validity data was reported.

A factor analysis of a battery of tests of perceptual and motor skills, based on the scores of first and fourth year dental students was reported by Zullo. Three dexterity factors, manual, finger and tweezer, and a spatial relations factor emerged in both groups. The dexterity factors were reasonably comparable across groups; the spatial factor was not.

B. Inventory Development and Validation

In this group of seven studies, three focused on pupil behavior, three on classroom activities, and one on teacher morale. Five of the reports are mainly descriptive of the instruments and their development. Two are concerned with validation.

Barclay described his *Classroom Climate Inventory*, which is designed to obtain self-judgments as peer judgments of a pupil's competencies, and the teacher's judgment of his adjustment, effort and motivation. Studies of

the reliability and validity of the instrument are referred to tersely, and its practical usefulness in the classroom is described.

The *Classroom Observational Record*, designed to aid observers in analyz- the cognitive levels of classroom verbal interactions, is described in a report by Reynolds and others. Their report discusses the usefulness of the record as a research tool and a training device.

Cassel and Pauk describe the development and standardization of *CLASSIC*, he *Cornell Learning and Study Skills Inventory*. Each of the seven part scores measures one of the seven factors identified as pertinent to learning problems: goal orientation, activity structure, scholarly skills, lecture mastery, textbook mastery, examination mastery, self mastery and study efficiency. Data on reliability and validity are provided, and uses of the inventory are suggested.

The *Pupil Behavior Inventory* is intended to help classroom teachers infer pupils' self concepts without relying on pupil self reports. In the revised form, developed by factor analysis from a preliminary form, the teacher records how frequently each pupil does each of eighteen different kinds of things. From these records, a score on relating, on asserting, on investing and on accomplishing is obtained for each pupil. Data shows the instrument to be reasonably reliable. Several kinds of validity evidence are also presented.

As part of a larger study of individual modes of coping with environmental demands and opportunities, Edwards developed a questionnaire to measure an adolescent's disposition to seek new experiences and to accept or promote social change. The questionnaire is shown to have good reliability and convergent validity when compared with peer ratings and semi-projective measures.

A study reported by Wahlstrom undertook to determine the value of the *Class Activities Questionnaire* as an instrument for describing the "environment" of a high school classroom. Factor analysis of extensive tryout data revealed eight principal factors in the 27-item questionnaire. Since scores for some of the 16 logical factors covered by the instrument showed rather low reliability coefficients, the investigator concluded that more items would be desirable. He also reported that many students and teachers find a *CAQ* threatening.

A canonical correlation analysis of two standardized measures of teacher morale revealed four interpretable variates: supportive relations, pay and benefits, work load, and facilities and equipment. The measures analyzed in this study by Coughlan and Froemel were the *Purdue Teacher Opinionaire* and the *School Survey*. The relation between *PTO* and *SS* turned out to be more complex than expected but the four principal variates correlate sufficiently to warrant the use of these instruments as alternate forms for measuring those four variates.

C. Measurement of Creativity

Of the five studies in this group which focused on creativity, two were concerned with the nature of creativity, two with tests of creativity, and one with the improvement of creativity test scores.

Using seven relatively pure measures of abilities expected to contribute to semantic creativity, Unks and Merrifield studied the stability of the factors across different grade levels and different socio-economic communities. Similar factor structures were found in the subgroups and in the total sample. Three structures of intellect factors in the area of divergent production were found. A fourth factor representing general language facility was defined by verbal I.Q. and strongly related to community.

Noting that Guilford in his studies of creative thinking focused on a limited (scientific) manifestation of creative behavior, and that Jackson and Messick have provided a broader framework in which unusualness, appropriateness, transformation and condensation provide the criteria, Feldman and others examined results from the Torrance tests in relation to the broader conception. They concluded that the Torrance tests do not tend to elicit high-powered creative responses.

Lynch reported on the development of a remote associates test for children in Grades 1-3. When the doublet cob-pop is presented the creative child is expected to respond with the word corn. Each form of the test, called the *Mini Rat* includes 20 doublet items. Analysis of results from administration of the test provide substantial support for its validity.

Greene reported encouraging results in the use of computerized content analysis of student responses to *Torrance Tests of Creative Thinking*. Reliability estimates were obtained from analysis of variance, and multiple

regression was used to maximize prediction of each subject's scores. The data analyses were detailed and sophisticated. However this report does not indicate how the computer was used to generate the data.

The effectiveness of Crutchfield's productive thinking program in increasing the divergent and convergent productive thinking abilities of sixth grade students was studied by Sporborg. He used two of Guilford's tests to measure divergent thinking and two to measure convergent thinking. The productive thinking program did not prove to be effective in this situation.

D. Factors in Test Performance

Seven studies in this collection had to do with factors in test performance. One dealt with a training program before the testing, two studied the effects of variations in study behavior, two others involved conditions within the examinee or in the testing situation, one investigated the effect of changing answers to objective test questions, and one the effect of knowledge of results.

The effects of four different perceptual training programs (general readiness, visual alphabet perception, auditory alphabet perception, and combined visual and auditory) on intelligence and reading readiness test scores were studied by Segal. The subjects were 54 disadvantaged 5 year olds in an OEO day care center. Score gains after 35 days on both intelligence and readiness tests were substantial with all programs, but the combined visual-auditory program appeared best over all.

Biggs factor analyzed scores on five measures used to determine grades in two college level educational psychology courses. The five measures were objective midterm and objective final tests, essay final, short answer final, and term paper. No reliabilities for these scores are reported, but the correlation between objective midterm and objective final is a surprisingly low .385. Three factors, general achievement, essay style and objective style were extracted. These were then correlated with scores obtained from a study behavior questionnaire. The investigator concluded that one method of evaluation did not necessarily promote one kind of study behavior as opposed to another.

The highlighting of particular statements in a text was found by Leicht and Cashen not to be appreciably effective in promoting retention. A set of four reprints was given to each of 164 college students, who were divided into four groups. In the reprints for three groups, three different kinds of statements were underlined: principles, examples and trivia. Reprints for the fourth group had nothing underlined. Questions on the underlined material were included in the regular class examinations. Scores on these questions were remarkably similar across groups.

In a study by Bennett and Entin 36 male educational psychology students filled out two questionnaires, one on test anxiety and another on long-term involvement, immediately after completing their final examination. The hypothesis that high anxious, highly motivated subjects would persist longer on the examination than low anxious, highly motivated subjects was supported. The hypothesis that higher test scores would go to those who perceive the course to have stronger future implications was not supported.

Group testing and giving of grades was found to result in higher scores on a creativity test than individual testing and absence of grades, in a study by Edwards. The influence of grades was stronger than the group influence. Subjects of the study were 131 sixth grade students. Creativity was measured by one item from the *Uses Test*.

Jacobs studied the effects of item difficulty and examinee ability on answer-changing behavior. Multiple-choice items were first presented on slides, for predetermined, fixed time periods. Then examinees received a printed copy of the test and a colored pencil, and were allowed to change any answers they wished without erasing the original answer. Fewest answers were changed on the easiest items, as might be expected. Changes improved examinee scores more often on easier than on more difficult items. No effects attributable to subject ability were observed. On the whole, the changes improved the examinee's score.

Beeson found evidence that immediate, item-by-item knowledge of results during the taking of multiple-choice tests did not depress, but slightly improved test performance. The subjects were 75 high school and college students enrolled in three mathematics classes. They received immediate knowledge of results from IBM card punchboards on half the items in the tests. For the other half of the items, knowledge of results was delayed.

E. Use of Tests to Measure Status or Change

Six studies fall in this category. Two deal with cognitive development, two with attitudes, and two with pupil behavior.

In a study of urban school children in six countries--Brazil, England, Italy, Japan, Mexico and Yugoslavia, Peck found that higher status children scored higher than lower status children on aptitude tests, achievement tests, and school grades. There is a notable lack of systematic sex differences in performance.

The ability of children in Grades 1-3 to apply the concepts of more and less to bipolar terms like high-low was studied by Harasym and others using the semantic differential technique. Logical conservers (Piaget) seem better able to apply more and less properly than do either intuitive conservers or non-conservers.

Nelsen and Johnson used the *College Student Questionnaire* to study attitude changes of students enrolled in predominantly black colleges and universities in their first year or two. Results showed general increases in Cultural Sophistication, and some increases in Family Independence, Peer Independence and Liberalism.

The semantic differential approach was utilized by Askov to build an instrument to measure teacher attitudes toward individualized instruction. A school situation and action in response to it are described in a brief introductory paragraph. Then several pairs of bipolar adjectives are provided on which teachers can react to the suggested action. The instrument was shown to have high reliability and significant validity.

The extent to which children repeat errors on repeated trials with the *Porteus Maze Test* was used by Burleigh and others to obtain a score which differentiates hyperkinetic children from normals, and which shows the beneficial effect of Ritalin. The study confirmed the belief that children with hyperkinesis do tend to repeat inappropriate behavior patterns more frequently than do normals.

Barclay showed that students nominated by their teachers and peers as most disruptive were characterized by teachers as restless, anxious and distractible. Those nominated as most reticent were characterized as introverted, cautious and controlled. He used analysis of variance and stepwise regression to identify scales whose scores contributed most to prediction of disruptive behavior or reticence.

F. Use of Tests to Predict

The four studies summarized in this section were concerned with the prediction of reading achievement in Grade 1, the prediction of freshman GPA, the prediction of success and satisfaction in a vocational program, and the prediction of teacher behavior.

A study of Harkham and others investigated the value of measures readily obtainable in kindergarten as predictors of reading achievement in Grade 1. *The Metropolitan Readiness Test* was best. The *Draw-a-Man Test* was poor. A behavior rating scale, and teacher rankings yielded correlations in the .40's and .50's. Inclusion of other measures with the *Metropolitan* improved predictions only slightly.

An *Admission Index*, derived from three sets of counselor ratings of academic and personal promise and a motivation scale, was found by Nicholson to be modestly correlated with freshman GPA, and to differentiate significantly between those who graduate with honors, graduate without honors, or do not graduate four years later.

Prediger related scores of 1600 prospective vocational school students on 36 aptitude, interest and personality measures to subsequent success and satisfaction. Discriminant analysis and regression analysis were made with computer assistance. He concluded that these procedures were highly effective in converting test data to counseling information.

Schluck studied linear, multiple and curvilinear relations between *MMPI* scores and data obtained from two classroom observation and record devices. She concluded that the *MMPI* might be useful in predicting future teacher behavior. She also noted that it is difficult to collect sufficient data in observational studies to arrive at dependable conclusions.

G. Use of Tests to Foster Learning

Three studies dealt with the use of tests to foster learning. One was concerned with the development of inferential thinking. Another used tests to develop skill in self appraisal. The third described procedures used in the study of medical thinking.

By giving weekly quizzes requiring Grade 8 history students to draw inferences, McKenzie improved their ability to draw inferences in history. This improvement however did not seem to transfer to unfamiliar subject matter.

Egelston asked students to predict what percentage score they would receive on each unit test both just before and just after its administration. The number of tests given to each class ranged from eight to thirteen. Some students, particularly those of higher ability, were able to improve their predictions during the course of the study.

The problem solving procedures of experienced physicians as they performed diagnotic work in a simulated medical setting were studied by Elstein and Shulman. They obtained records of the thinking behavior by asking the physician to think aloud, by retrospection during the work up, and by video-tape-simulated retrospection. They observed that efficiency in diagnosis required generation of hypotheses which are strong conceptual competitors.

H. Concluding Statement

Review of these 39 reports left the reviewer with several general impressions. The dominant one is that many capable workers are doing a great deal of good research work. The level of competence exhibited in experimental design and statistical analysis is adequate to high. In a few instances it appears that the design is over elaborate in relation to the problem, and that the analysis is over extended so that, in effect, a razor is being used to cut butter.

On the whole the level of reporting is good too, direct, concrete, succinct and well organized. In only a few cases is the reader required to reconstruct from clues scattered about by the writer what particular problem he was attacking, how he proceeded to attack it, and what he found. A few writers have not learned that a plethora of tables, with massive arrays of data in each table, are more likely to conceal than to reveal the message they are trying to convey. Discursive discussions of background, related research and implications of the findings can have a similar effect. Succinctness is always a virtue, but never more so than in a report of research to be presented at an annual meeting.

Despite occasional shortcomings, the quality of these reports was high. They make substantial contributions to our understanding of the processes of education, and they point to areas where continuing investigation is likely to be fruitful.

List of Papers Reviewed

Askov, E. N. Assessment of teachers' attitudes toward an individualized approach to reading instruction. 19p. (ED 048 349; MF and HC available from EDRS).

Barclay, J. R. Characteristics of reticent and disruptive children as identified by the *Barclay Classroom Climate Inventory*. From symposium "Measuring the classroom climate." 17p. (ED 051 278; MF and HC available from EDRS).

Barclay, J. R. Measuring the social climate of the classroom. From symposium "Measuring the social climate of the classroom." 12p. (ED 051 277; MF and HC available from EDRS).

Beeson, R. O. Immediate knowledge of results and test performance. 8p. (ED 048 375; MF only available from EDRS).

Bennett, C. R., & Entin, E. E. The effects of test anxiety, course importance, and future orientation on persistence and academic performance. 9p. (ED 048 344; MF and HC available from EDRS).

Bierly, M. M. A validation of a method of assessing young children's language competence. 16p. (ED 051 289; MF and HC available from EDRS).

Biggs, J. B. Effects of study behavior on objective-style and essay-style performance. 50p. (ED 048 350; MF and HC available from EDRS).

Burleigh, A. C., & Others. Development of a score that separates hyperkinetic and normal children and demonstrates drug effect. 9p. (ED 048 374; MF and HC available from EDRS).

Cassel, R. N., & Pauk, W. J. Development and standardization of the *Cornell Learning and Study Skills Inventory (CLAS-SIC)*. 10p. (ED 048 361; MF and HC available from EDRS).

Coughlan, R. J., & Froemel, E. C. A comparison between two standardized measures of teacher morale. 14p. (ED 050 170; MF and HC available from EDRS).

Ebel, R. L. The comparative effectiveness of true-false and multiple choice achievement test items. 5p. (ED 050 148; MF only available from EDRS).

Edwards, D. W. The development of a questionnaire method of measuring exploration preferences. From symposium "Copying styles and the High School." 9p. (ED 051 285; MF and HC available from EDRS).

Edwards, T. M. The effects of environment on performance during creativity testing. 11p. (ED 048 376; MF and HC available from EDRS).

Egelston, R. L. Test achievement: Expectation and reality. 21p. (ED 049 273; MF and HC available from EDRS).

Elstein, A. S., & Shulman, L. S. A method for the study of medical thinking and problem solving. 31p. (ED 050 164; MF and HC available from EDRS).

Feldman, D. H., & Others. Unusualness, appropriateness, transformation and condensation as criteria for creativity. 18p. (ED 050 166; MF and HC available from EDRS).

Fischbach, T. J. Study of relationships of reading mastery level to general reading achievement to validate diagnostic reading tests. 24p. (ED 049 285; MF and HC available from EDRS).

Greene, J. F. Computer simulation of human behavior: Assessment of creativity. From symposium "Multiple regression prediction models in the behavioral sciences." 17p. (ED 049 292; MF and HC available from (EDRS).

Harasym, C. R., & Others. Use of "more" and "less" in conservation: A semantic differential analysis. 11p. (ED 046 980; MF and HC available from EDRS).

Harckham, L. D., & Others. Multiple prediction of reading achievement in grades one through four using kindergarten measures. 11p. (ED 049 311; MF and HC available from EDRS).

Hedl, J. J., Jr., & Others. Computer-based intelligence testing. 22p. (ED 050 146; MF and HC available from EDRS).

Jacobs, S. S. An experimental analysis of answer-changing behavior on objective tests. 10p. (ED 048 345; MF and HC available from EDRS).

Leicht, K. L., & Cashen, V. M. Type of highlighted material and examination performance. 9p. (ED 050 155; MF and HC available from EDRS).

Lynch, M. D. The *Mini Rat*: Its development and some evidence on its validity. 17p. (ED 050 149; MF and HC vailable from EDRS).

McKenzie, G. R. Facilitating inferential thinking with weekly quizzes. 6p. (ED 046 996; MF and HC available from EDRS).

Nelsen, E. A., & Johnson, N. C. Attitude changes on the *College Student Questionnaires*: A study of students enrolled in predominantly black colleges and universities. 33p. (ED 049 296; MF and HC available from EDRS).

Nicholson, E. The *Admission Index* as a predictor of freshman GPA. 12p. (ED 049 283; MF and HC available from EDRS).

Peck, R. F. A cross-national comparison of sex and socio-economic differences in aptitude and achievement. 13p. (ED 049 315; MF and HC available from EDRS).

Prediger, D. J. Converting test data to counseling information system implementation in a vocational school. 51p. (ED 050 161; MF and HC available from EDRS).

Purkey, W. W., & Others. The development of a *Pupil Behavior Inventory* to infer learner self concept. 11p. (ED 050 153; MF and HC available from EDRS).

Reynolds, W. W., Jr., & Others. *The Classroom Observational Record.* 15p. (ED 048 378; MF and HC available from EDRS).

Schluck, C. Using the *MMPI* to predict teacher behavior. 10p. (ED 049 313; MF and HC available from EDRS).

Scott, N. C. Cognitive style assessment: One test or several? 9p. (ED 048 355; MF and HC available from EDRS).

Segal, M. Effects of four different perceptual training programs on IQ and reading readiness in the lower socio-economic level kindergarten child. 74p. (ED 046 974; MF and HC available from EDRS).

Sporborg, A. The effect of programmed productive thinking materials on the divergent and convergent test scores of sixth grade students. 17p. (ED 049 284; MF and HC available from EDRS).

Unks, N. J., & Merrifield, P. R. Stability of productive thinking factors across different communities and grade levels. 18p. (ED 049 274; MF and HC available from EDRS).

Wahlstrom, M. W. Factorial validation of the *Class Activities Questionnaire.* 26p. (ED 051 276; MF and HC available from EDRS).

Whalen, T. E. A validation of the Smith test for measuring teacher judgment of written composition. 7p. (ED 049 275; MF and HC available from EDRS).

Zullo, T. G. Patterns of perceptual motor skills in first and fourth year dental students. 10p. (ED 044 445; MF and HC available from EDRS).