DOCUMENT RESUME

ED 073 141 TM 002 387

AUTHOR	Ebel, Robert L.
TITLE	Test Development, Interpretation, and Use.
INSTITUTION	Educational Testing Service, Princeton, N.J.; ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	ERIC-TM-19
PUB DATE	Feb 73
NOTE	10p.; 1972 AERA Conference Summaries
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Conference Reports; *Measurement Instruments; *Measurement Techniques; Speeches; *Test Construction; *Test Interpretation; *Test Reviews

ABSTRACT

The 54 papers related to test development, interpretation, and use that were presented at the 1972 AERA Conference are reviewed. The papers were classified into 11 categories, as follows: A. What to measure--educational objectives; attitude measurement; and creativity; B. How to measure--item types; test development; response modifications; confidence weighting; semantic differential and observational techniques; and C. Test use--testing programs; and test bias. A listing of the papers reviewed, their authors, and, when applicable, the ED numbers concludes the summary. (DB)

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

TM Report 19

February 1973

1972 AERA Conference Summaries

TEST DEVELOPMENT, INTERPRETATION, AND USE

Robert L. Ebel

Michigan State University

This publication was prepared pursuant to a contract with the National Institute of Education, U. S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy.

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
   ED 058 274

6. State Educational Assessment Programs: An Overview
   ED 058 309

7. Criterion Referenced Measurement. A Bibliography
   ED 060 041

8. Measures Pertaining to Health Education: I. Smoking
   ED 060 042

9. Measures Pertaining to Health Education: II. Drugs. An Annotated Bibliography
   TM 002 078 (ED Number not yet available)

10. Measures Pertaining to Health Education: III. Alcohol. An Annotated Bibliography
    TM 002 079 (ED Number not yet available)

11. 1971 AERA Conference Summary
    I. Evaluation: State of the Art
    ED 060 043

12. 1971 AERA Conference Summary
    II. Criterion Referenced Measurement
    ED 060 134

13. 1971 AERA Conference Summary
    III. Educational Statistics
    ED 060 133

14. 1971 AERA Conference Summary
    IV. Test Development, Interpretation, and Use
    ED 060 135

15. 1971 AERA Conference Summary
    V. Innovations in Measurement
    ED 060 044

INTRODUCTION

About 700 of the 1,000 papers presented at the 1972 AERA Annual Meeting in Chicago, Illinois 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)* 200 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. Four topics were chosen for the series: Criterion Referenced Measurement, Evaluation, Statistics, and Test Construction.

Most 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 0
Bethesda, Maryland 20014

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

TEST DEVELOPMENT, INTERPRETATION, AND USE

Robert L. Ebel

The 54 papers in this area which were presented at the annual meeting in Chicago, April 1972, were classified into eleven categories as shown below. The figure in parenthesis is the number of studies in the category.

A. What to measure
   1. Educational objectives (3)
   2. Attitude measurement (7)
   3. Creativity (7)

B. How to measure
   4. Item types (2)
   5. Test development (7)
   6. Response modifications (5)
   7. Confidence weighting (4)
   8. Semantic differential (5)
   9. Observational techniques (3)

C. Test use
   10. Testing programs (8)
   11. Test bias (3)

1. Educational Objectives

Hoepfner obtained ratings of 106 goal statements from 2,555 individual raters. Self esteem received the highest rating (4.67 on a five-point scale of importance). Writing fluency in a foreign language received the lowest rating (1.48). The author cautions against the assumption of general validity for his findings, but believes that the procedure used is worthy of wider application.

A Q-sort was used by Doherty to differentiate the goal valuations of elementary school teachers, principals and parents when classified along demographic lines such as racial-ethnic composition of the student body He found that the demographic variables seemed to have little influence on the priority rating of the goals.

Dyer and others describe a methodologically sophisticated utility function estimation procedure designed to provide curriculum planning information to elementary school principles. The reader who is not already well versed in utility function methodology may be inclined to question the authors' claim that it is a "simple procedure." An elementary school principal may have difficulty in perceiving its utility for him.

2. Attitude Measurement

Kilbane reported development via factor analysis of a 30 item inventory designed to tap high school students' attitudes toward self, toward teachers and toward social participation. She reported that the instrument appeared to differentiate between students who remained in school and those who dropped out. A similar effort to develop an instrument to measure the attitudes of seventh grade pupils toward school learning was reported by Brehman. The new instrument was shown to have high internal consistency. No validity data was reported.

Assessment of the quality of education in the public schools of Pennsylvania has been a high priority objective of the Pennsylvania Department of Education in recent years. Brehman's study and the two which follow were outgrowths of efforts to assess student attitudes toward school learning.

McGuiness and Stanl: used the technique of sociometry to validate a measure of understanding and appreciation of persons belonging to social, cultural and ethnic groups different from one's own. The instrument showed high (.95) internal consistency, moderate (.79) retest reliability after seven months, and low to moderate validity coefficients (.28 to .73).

Landis reported a study of the construct validity of a self esteem inventory. It was predicted that students making higher scores on a standardized achievement test battery would also make higher scores on the self esteem inventory. It was also predicted that scores on the self esteem inventory would correlate positively with scores on a self concept as a learner scale. Both predictions were substantiated.

The influence of a mothers self concept on the deprived child's self concept has been investigated in several studies by Lecco and Bridges. The present report cites several significant relationships which were demonstrated by canonical correlation techniques

Gable and Pruzek compared the results of two methods of categorizing the items in a scale for measuring attitudes toward black people. One method applied latent partition analysis to classifications of the items by judges. The other method applied factor analysis to the responses of college students to the same items. The two methods showed substantial agreement. The use of this combination of approaches provides some validation of the attitude constructs being assessed by the instrument.

In a brief report Shoemaker described a method for reducing the labor of scaling attitude test items by the method of paired comparison. Randomly selected subsets of pairings were given to random samples of the judges. He concluded that the sampling procedure yielded scale values satisfactorily approximating those obtained from the whole population of pairings and of judges. No data

of results of data analysis are included in the report, however.

3. Creativity

A 29 item inventory of creative accomplishment was administered to 166 college students, along with three tests of convergent thinking and five of divergent thinking, in a study reported by Stafford and Browne. The data indicated to the authors that creative accomplishment is related more closely to fluency than it is to originality, convergent thinking or drive. The data, however, were not included in the report.

A study of the relation of three cognitive styles, response tempo, response style and response ambiguity, to creative problem solving was reported by Hyer and Rookey. Analysis of data resulting from administration of tests of intelligence, cognitive style and creativity to 288 junior high school students led to the conclusion that creative problem solving is about equally affected by intelligence and the cognitive styles.

Using as criteria the ratings by two professors of the creativity of 34 graduate student writers, Barro found higher correlations with personality than with cognitive test scores. The correlations tended to be low, and not significantly different from zero for a majority of the possible predictor variables. No reliability coefficients were reported for any of the variables.

Stallings and Gillmore investigated the relation between measures of creativity and grades in courses presumably eliciting creative behavior. Measures of creativity were obtained from the Torrance Figural Test for over 300 freshmen in the College of Fine and Applied Arts at the University of Illinois. Only one validity coefficient, of the 68 generated, was significant at the .05 level. The authors concluded that scores from this test had little utility in enhancing the prediction of grades in these courses, and that their data did not support the validity of the test.

The effects of practice on the nonverbal creativity of fifth grade children was investigated by Roweton and Spencer, using forms A and B of Torrance's picture completion task. They found the overall effects to be less marked than in the case of verbal creativity, and more dependent on peculiarities of the task item.

Shigaki also studied the effect of practice (trend of scores over time) on originality scores obtained from 56 protocols resulting from administration of the Torrance Tests of Creative Thinking to intermediate grade children. She found significant improvement in verbal originality test scores but not in the figural test scores.

Prediction equations have been developed to simulate by computer the behavior of human judges in rating creativity test performances. Greene and Zirkel undertook to determine the stability and usefulness of these equations when applied to samples drawn from other populations. They concluded that equations predicting ratings of fluency and flexibility were stable enough to be useful, but that those designed to predict ratings of originality were not.

4. Item Types

In a study comparing multiple-choice and true-false test items, Oosterhof and Glasnapp found lower reliability for tests composed of true-false items than for tests composed of multiple choice items even after adjustment for differences in time requirements. They also found that false versions of an item contributed more to reliability than true versions, and that multiple choice items required about 1.75 times as much time to answer as true-false items. Subjects for their study were 101 undergraduates enrolled in an introductory measurements course.

In a similar study using 1,018 high school students as subjects, Frisbie and Eble obtained similar results. Their subjects required 1.5 times as much time to answer a multiple choice item as they required to answer a true-false item. In all cases the true-false tests were less reliable than the multiple choice tests even after adjustment for differences in time required. The data from this study did not justify rejection of the hypothesis that the two types of tests measure the same thing.

5. Test Development

Orpet's report on the development of an experimental sensory motor and movement skills test battery emphasizes the potential usefulness of such a battery, provides a brief rationale for each of the tests included in it, describes the standardization procedure, and gives lower bound reliability estimates derived from communalities from the factor analysis. No data on the score distributions from the various tests for children in various grades are reported.

Pandey and Cleary describe the development of a test of basic skills for adults enrolled in literacy and other remedial programs. The test includes 20 communications items, 19 numerical items and 28 items measuring practical skills in such things as using a telephone directory or filling out forms. The subtests are shown to have high reliability and to differentiate seventh, eighth and ninth grade students clearly.

Development of a test to measure occupational awareness was described in a report by Reardon and others. The items were based on information from the dictionary of occupational titles with emphasis on worker traits in different occupations. No sample items are included in the report. The final form of the test had a KR 20

reliability of .769 in statewide administration to 2,640 pupils in 90 schools in Pennsylvania. (It is interesting to note that the number of items in the test is reported to be 30.000!) School means ranged widely from 10.57 to 21.77, indicating that the test will discriminate clearly among schools.

Millman discussed several bases for determining the passing score on a criterion-referenced test. He also described a means for determining the relation between the passing score on a test, the number of items in the test, and the percent of students wrongly passed and failed. A table illustrating these relationships is included in the report.

A computer program for textual analysis was used by Felsenthal and Felsenthal to obtain data from 20 trade books for children for computation of readability indices by various formulas. Data on the means, standard deviations and intercorrelations of the various readability measures is reported. Analysis of variance showed no significant difference in the readability of material in the first, middle or last thirds of the books. No data are presented on the advantages or disadvantages of computer assistance with this task.

Hofmann, in a long and elaborately mathematical paper discussed an efficiency index for use in item analysis. It is defined as the ratio of the observed discrimination to the maximum possible discrimination for an item of that difficulty level. The author shows that the index lends itself to a variety of interpretations, many of which can be given as probability statements. He suggests, but does not demonstrate, that the use of this index will enable test-makers to build better tests.

In a long report of a complex study Tyler analyzed the relation of response stability to personality test homogeniety, and reported data gathered to test the analysis. The data were obtained from 22 dichotomously scored personality tests. Persons and items were scaled according to the Rasch model. Tyler found that response instability was a joint function of subject and item scale values. While heterogeneous tests provide the best predictions in practical applications, homogeneous tests have more to contribute in the development of personality theory.

6. Response modifications

Koehler reviewed the rationale for and experimental assessments of several modifications of conventional best answer response to multiple choice test items. One of these asks the examinee to cross out every response he knows to be incorrect. Another asks him to work as many as necessary to be sure of including the correct response. Koehler concluded that the experimental data provide substantial evidence in favor of the continued use of the conventional one-best-answer response to multiple choice test items.

Reilly and Jackson reported that empirical response weighting of the multiple choice items in the aptitude test of the Graduate Record Examination substantially increased test reliability but did not increase test validity. They suggested that the empirical weighting capitalized on the tendency to omit, and that while this tendency is reliable, it is not valid.

Scores obtained from exact and approximate Guttman weights were compared by Green with several other weighted and unweighted scores for 2,500 men on the verbal portion of the Scholastic Aptitude Test. He concluded that weighted scoring is not to be recommended in this situation.

Hendrickson and Green studied the effect of Guttman weighted scoring on the factor structure of subtests of the Scholastic Aptitude Test, using rights-only scoring as the basis for comparison. They found significant differences in the factor structure of the two types of scores and concluded that the two scores measure different functions. This helps to explain why Guttman-weighting which increases the reliability of test scores often reduces their validity.

Baker showed that the item response weighting technique known as the method of reciprocal averages is a particular implementation of Guttman's general model for internal consistency scaling. He argued that the reciprocal averages method is computationally simpler than the Guttman method, and that it can be implemented by a simple extension of existing item analysis computer programs.

7. Confidence weighting

Twenty-four graduate education majors enrolled in a course on measurement and evaluation took a 20 item test on which they had the option of either confidence weighted response or Coombs-type multiple response. Garvin, who conducted and reported the study, found that the weighted scoring procedures separately or in combination invariably depressed the reliability coefficient.

A second paper by Garvin presented a comprehensive discussion of confidence weighting; its nature, purpose, varieties and effects. It is his opinion that confidence weighting procedures have a kind of intrinsic validity, for knowing how much confidence to place in a belief is an important aspect of a person's knowledge.

In two closely related papers, Rippey discussed the rationale and development of confidence testing. He made a case for the use of "intrinsic items" which do not have unique correct responses and require the examinee to distribute his belief over the options. Several mathematical functions that might be used in scoring confidence

weighted responses were suggested. He concluded that effective use of confidence testing will require the solution of a number of psychological and educational problems.

8. The Semantic Differential

Raper and Wasik used the semantic differential technique to differentiate pre-school educational environments. They found that perceptions of the pre-school environment were related to the prior experience of the perceiver, and to the conditions under which the semantic differential data were gathered.

Two hundred sixteen elementary school children rated the concept "myself" on semantic differential scales defined by 55 adjective pairs in a study reported by Lynch and Cochran. Ratings of the 55 scales were factor analyzed overall and by grade. Three factors were extracted overall, but six were found in grade two and seven in each of grades four and six. This is a more complex judgmental structure than has previously been reported for grade school children.

The semantic differential technique was used by Francies to measure the attitudes of sixth grade pupils toward a course in Family Life education. Drawings depicting family situations were displayed to the pupils by means of an overhead projector. The investigator concluded that this technique of attitude measurement merits wider use.

Gulo summarized the results of four recent studies involving use of the semantic differential to measure teaching effectiveness. The factors revealed in these four studies differ somewhat Those that reappear tend to account for different proportions of the total variance. Nevertheless the author regards the technique as especially useful in quantifying student perceptions of effective teaching.

In an essentially methodological paper, Lynch urged more frequent use of the D statistic (generalized distance function) as a basis for comparing profiles on semantic differential data. He suggested that this approach would facilitate interpretable results on meaningful variables with both efficiency and theoretical utility.

9. Observational Techniques

Cunningham and Boger described the Parent-Child Interaction Rating Procedure. Video-taped sessions in which the parent teaches the child to perform a simple sorting task provide a record of the interactions which are rated. Among the aspects of interaction that are rated are voice tone, task orientation, reward and cues. The authors of the report see wide usefulness of the procedure in the study of the teaching-learning of young children, and the development of their behavior patterns.

The construction and validation of a theoretically based system for the analysis of teaching roles in childhood education was reported by Southwell and Webb. Four teaching roles—acquisition, inquiry, mother-surrogate, and authenticity—were defined, and teaching behaviors presumably characteristic of each role were identified. The investigators found that teachers in different schools and in different educational programs exhibited characteristically different behaviors.

The rod and frame test was administered twice to 70 children, 5 to 7 years of age, along with an intelligence test Significant differences were found for age groups and sex groups, but not between first and second test administrations. Reliability coefficients ranged from .43 to .72 for different age groups and with different intervals between testing. Correlation of rod and frame test scores with intelligence test scores was low. The investigators, Busch and Simon, conclude that further definition of the construct will be necessary.

10. Testing Programs

Seven papers in this group were presented as part of a symposium on "The Madison Plan: A New Approach to System-Wide Testing." Mathews, stated the purpose of the symposium and described its structure. The symposium, he said, takes a long hard look at a school district testing program, finds it inadequate and dysfunctional, proposes an alternative structure, and discusses the results from, and problems with, attempts to implement this structure.

Cleary and Mathews described how dissatisfaction with the existing program, which involved massive testing with minimum use of the results, led to organization of the Nucleus Testing Committee. This committee representing all schools in the system was charged with becoming knowledgeable about tests and determining what kind of data the schools needed about children.

Seeman reported results of a survey of the evaluation concerns of various members of the school staff. The survey revealed continuing concern for the capacity/achievement dimension, and some discordant priorities of various staff groups and staff members.

How and why the testing program of the Madison Public Schools was reduced to reading in grades 1,2, 3, 4, 5, and 8, and mathematics in grade 5 was described by Nettleton. She noted that when more specific criterion-referenced tests are adopted, testing every pupil will be replaced by random sampling to provide the same normative data for less time and cost.

Christiansen described the work of the Curriculum-Related Subcommittee, its accomplishments and plans for further development of curriculum-related tests. The subcommittee is emphasizing objective-based instruction,

criterion-referenced testing and program-fair assessment.

One facet of the "Madison Plan" for system wide testing called for exploration of testing in the affective domain. A report by Hansen describes what the affective subcommittee did during 18 months to acquire an understanding of the affective domain and to develop criteria for an affective testing program. Creation of the tests remains a task for the future which, in the view of the committee, is brighter.

Presenting an administration view of the "Madison Plan", Sapone (Director of Curriculum Development in Madison Public Schools) rejects the assumptions of norm referenced testing, deplores its effects on teaching and learning, and suggests that it ought to disappear quickly. He believes that the new criterion referenced testing program being developed in Madison should begin to pay dividends within the next few years not only to Madison but to the whole nation.

In another paper Mathews described computer generated verbal reports of test scores for parents and teachers. These reports supplement numerical reports and eliminate the need for tables of percentiles or grade equivalent scores. Teachers tended to rate the verbal reports higher than the conventional reports in clarity, usefulness, meaning, value, sufficiency and accuracy.

11. Test Bias

Do eighth grade students from minority groups perform differently on tests of academic aptitude and achievement when they know their scores will be compared with scores of (1) other minority students or (2) majority group students? The answer from a study by Oakland and Emmer is negative. However the nature of the comparison group did have some effect on their expectations of performance level.

Is the performance of a first, second or third grade pupil on an individual test of intelligence affected by the race of the examiner? Savage and Bowers found the answer to be yes in their study. Pending further study of the complex interactions they recommend that tester and student tested be of the same race.

Green investigated the possibility that an elementary school achievement test might be biased against minority group members. In his view a test is biased against a particular group if it contains a substantial proportion of items that would not have been selected if the item tryout had been made in that particular group. He found substantial evidence for bias of that kind, and suggested that producers of standardized tests would show more concern for eliminating it.

12. Concluding Remarks

On the whole these reports reflect competently executed research studies. In a few cases essential data that should have been available to the researcher were not given in the report, and in some cases the reader is left in doubt concerning what particular questions the researcher was trying to answer and what answers he thought he had found. Critical readers, basing reactions on their own special interests and perceptions of truth, are likely to see shortcomings in many of the studies reported. But they are also likely to learn something of value from most of them.

PAPERS REVIEWED

Baker, F.B., & Hoyt, J. The relation of the method of reciprocal averages to Guttman's internal consistency scaling model. 19p. (ED 062 397, MF and HC available from EDRS.)

Barro, A.R. Personality and cognitive correlates of creativity in writers. 14p. (ED 064 394, MF and HC available from EDRS.)

Brehman, G.E.,Jr. Attitude toward school learning: The development of a seventh grade level instrument for measurement of goal IV of the Pennsylvania educational quality assessment program. 21p. (ED 062 391, MF and HC available from EDRS.)

Busch, J.C., & Simon, L.H. Methodological variables in the study of field dependent behavior of young children. 15p. (ED 063 328, MF and HC available from EDRS.)

Christiansen, P. Curriculum-related testing: An improvement program. 9p. (ED 064 338, MF and HC available from EDRS.)

Cleary, T.A., & Mathews, W.M. The Madison plan: A new approach to system wide testing. 9p. (ED 064 335, MF and HC available from EDRS.)

Cunningham, J.L., & Boger, R.P. Development of an observational procedure for assessment of parent-child interaction. 34p. (ED 064 320, MF and HC available from EDRS.)

Doherty, W.J. Differential valuations of elementary educational goals. 25p. (Document not yet available from EDRS.)

Dyer, J.S., & Others. Utility functions for test performance. 33p. (ED 064 330, MF and HC available from EDRS.)

Felsenthal, N.A., & Felsenthal, H. Utilizing the computer to assess the readability of language samples. 11p. (ED 061 021, MF and HC available from EDRS.)

Francies, H. Attitude measurement of pupils with varying reading abilities semantic differential using concepts presented by transparencies. 20p. (ED 064 382, MF and HC available from EDRS.)

Frisbie, D.A., & Ebel, R.L. Comparative reliabilities and validities of true-false and multiple choice tests. 8p. (ED 064 388, MF and HC available from EDRS.)

Gable, R.K., & Pruzek, R.M. Methodology for instrument validation: An application to attitude measurement. 26p. (ED 064 401, MF and HC available from EDRS.)

Garvin, A.D. Confidence weighting. 12p (ED 062 401, MF and HC available from EDRS.)

Garvin, A.D. Confidence weighting plus Coombs-type response options: A good idea that failed. 7p. (ED 065 551, MF and HC available from EDRS.)

Green, B.F. The sensitivity of Guttman weights. 8p. (ED 064 323, MF and HC available from EDRS.)

Green, D.R. Racial and ethnic bias in test construction. 30p. (ED 056 090, MF and HC available from EDRS.)

Greene, J.F., & Zirkel, P.A. Scoring creativity tests by computer simulation: A validation of prediction equations. 7p. (ED 062 403, MF and HC available from EDRS.)

Gulo, E.V. Measuring dimensions of teaching effectiveness with the semantic differential. 14p. (ED 064 346, MF and HC available from EDRS.)

Hansen, L.H. Toward a program for testing in the affective domain. 7p. (ED 064 337, MF and HC available from EDRS.)

Hendrickson, G.F., & Green, B.F. Comparison of the factor structure of Guttman weighted vs. rights-only-weighted tests. 14p. (ED 062 389, MF and HC available from EDRS.)

Hoepfner, R. National elementary education priorities. 8p. (Document not yet available from EDRS.)

Hofmann, R.J. The efficiency index in item analysis. 38p. (ED 064 355, MF and HC available from EDRS.)

Hyer, L., & Rookey, T.J. Cognitive style and creative problem solving. 13p. (ED 060 063, MF and HC available from EDRS.)

Kilbane, M.T. Development of an instrument to assess attitudes of high school students in compensatory programs. 11p. (ED 063 329, MF and HC available from EDRS.)

Koehler, R.A. Coombs' type response procedures. 13p. (ED 063 338, MF and HC available from EDRS.)

Landis, J.H. A validity study of the self-esteem inventory. 25p. (ED 062 392, MF and HC available from EDRS.)

Lynch, M D. Multidimensional measurement with the D statistic and the semantic differential. 18p. (ED 064 316, MF and HC available from EDRS.)

Lynch, M., & Cochran, T. Development and validation of a set of sematic differential scales for children. 19p. (ED 064 341, MF and HC available from EDRS)

Mathews, W.M. An alternative to a standardized testing program. 3p. (ED 064 334, MF and HC available from EDRS.)

Mathews, W.M. Computer-generated verbal testing reporting. 14p. (ED 064 358, MF and HC available from EDRS.)

McGuiness, T.P., & Stank, P.L. Model for use of sociometry to validate attitude measures. 8p. (ED 064 410, MF and HC available from EDRS.)

Millman, J. Passing scores and test lengths for domain-referenced measures. 17p. (ED 065 555, MF and HC available from EDRS.)

Nettleton, A.L Taming the standardized testing program. 13p. (ED 064 336, MF and HC available from EDRS.)

Oakland, T., & Emmer, E. Effects of knowledge of criterion groups on the test performance of Negro and Mexican-American students. 13p. (ED 064 379, MF and HC available from EDRS.)

Oosterhof, A.C., & Glasnapp, D.R. Comparative reliabilities of the multiple choice and true-false formats. 5p. (ED 064 361, MF and HC available from EDRS.)

Orpet, R.E. The development of an experimental sensory-motor and movement skills test battery. 9p. (ED 062 365, MF and HC available from EDRS.)

Pandey, T.N., & Cleary, T.A. The Wisconsin test of adult basic education. 12p. (ED 064 352, MF and HC available from EDRS.)

Raper, T.R., & Wasik, J.L. Use of the semantic differential in describing a pre-school environment. 12p. (ED 064 405, MF and HC available from EDRS.)

Reardon, F.J., & Others. The development and evaluation of a test to measure occupational awareness. 12p. (ED 064 371, MF and HC available from EDRS.)

Reilly, R.R., & Jackson, R. Effects of item option weighting on validity and reliability of shortened forms of the GRE aptitude tests. 14p. (ED 062 402, MF and HC available from EDRS.)

Rippey, R. An analysis of several effects of confidence testing. 14p. (Document not yet available from EDRS.)

Rippey, R.M. Scoring and analyzing confidence tests. 11p. (ED 060 070, MF and HC available from EDRS.)

Roweton, W.E., & Spencer, H.L., Jr. Facilitative effects of practice upon nonverbal creativity. 5p. (ED 059 563, MF and HC available from EDRS.)

Sapone, C.V. An administrative view. 19p. (ED 064 333, MF and HC available from EDRS.)

Savage, J.E , & Bowers, N.D. Testers' influence on children's intellectual performance. 8p. (ED 064 329, MF and HC available from EDRS.)

Seeman, M. Evaluation concerns of various members of the school staff. 5p. (ED 064 339, MF and HC available from EDRS.)

Shigaki, I.S. An analysis of the trend of originality scores on a measure of creativity. 9p. (ED 064 317, MF and HC available from EDRS.)

Shoemaker, D.M. An application of item-examinee sampling to scaling at [illegible] (ED 060 026, MF and HC available from EDRS.)

Southwell, R.K., & Webb, J.N. Development and validation of an observation system for analyzing teaching roles. 18p. (ED 064 368, MF and HC available from EDRS.)

Staffard, R.E., & Browne, W. Construct validity of creativity. 3p. (ED 064 384, MF and HC available from EDRS.)

Stallings, W.M., & Gillmore, G.M. Relationships between figural creativity and grades in a college of fine and applied arts. 11p. (ED 062 924, MF and HC available from EDRS.)

Tocco, T.S., & Bridges, C.M., Jr. A replication and an example of serendipity in educational research. 21p. (ED 062 396, MF and HC available from EDRS.)

Tyler, T.A. Test homogeneity and response stability. 39p. (ED 064 378, MF and HC available from EDRS.)