DOCUMENT RESUME

ED 060 044 TM 001 127

AUTHOR Swineford, Frances
TITLE 1971 AERA Conference Summaries: Innovations in Measurement.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
REPORT NO TM-R-15
PUB DATE Jan 72
NOTE 12p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Achievement Tests; Attitude Tests; Behavior Rating Scales; *Classroom Techniques; Computer Oriented Programs; *Conference Reports; Evaluation Techniques; *Innovation; *Measurement Instruments; *Measurement Techniques; Post Testing; Pretests; Rating Scales; Research Projects; Statistical Analysis; Statistics; Test Reliability; Test Validity; Weighted Scores
IDENTIFIERS *American Educational Research Association

ABSTRACT

Seventeen 1971 AERA presentations concerning new measurement procedures or new application of old procedures are summarized. Papers are grouped as follows: 1) statistics and measurements of interest to classroom teachers; 2) measures appropriate for use by a specialist; 3) research reports that increase knowledge of existing devices. (MS)

TM REPORTS

NUMBER 15

1971 AERA Conference Summaries

V. Innovations in Measurement

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.

January 1972

INNOVATIONS IN MEASUREMENT

Frances Swineford

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 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

Seventeen papers presented at the 1971 AERA Convention are reviewed here which fall within the general field of new measurement procedures or new applications of previously described procedures. As a result of the diversity of the topics represented, these papers could probably be classified in any of several ways. For present purposes three groupings have been used. The first concerns statistics and measurements that can be regarded as being within the realm of the classroom teacher rather than as serving theoretical statisticians. The second group deals with measures that would be expected to lie beyond the kind of application appropriate to a single classroom. The last group is a collection of papers whose principal contributions are analyses of measurement devices themselves as distinguished from applications in particular situations. Although no classification of statistical material of this nature can be expected to result in mutually exclusive groups, it is hoped that the present organization will prove to be a useful one.

Group A. Classroom. Papers in this group are deemed to be of interest to classroom teachers and others whose interests might be described as more practical than theoretical.

McMorris has collected three short-cut formulas for approximating a standard deviation, nine for approximating a Kuder-Richardson Formula 20 reliability coefficient, and two for approximating a standard error of measurement. He applied these to data for eighty-five teacher-made tests and shows which short-cuts provide the best approximations.

Peper and Chansky describe their scaling technique for sociometric data in terms of applications in twenty-two classes in arithmetic. Readily administered and scored, their device shows high reliability over a six-month period in the area in which it was used for this study.

Attacking the difficult problem of increasing the reliability of "change" scores or "growth" scores obtained from comparisons of pretest and posttest data, Crocker and Mehrens have experimented with four methods that produce more reliable change scores than are usually obtained. The methods were tested by the technique of cross validation.

Taylor and Helmstadter describe the construction and preliminary trial of a pair comparison test designed to measure aesthetic judgment of four- and five-year-old children. The scale consists of thirty-eight pairs of color slides picturing sculpture, paintings, household items, and the like. Not only was each slide judged on an 11-point scale of aesthetic quality by ten art experts from various fields of art, but each pair was also rated by the experts. Despite a notable lack of unanimity with respect to the scale judgments, there is no pair for which less than 70 per cent of the experts preferred the slide with the higher median scale value. The reliability of measures obtained for forty children tested twice, six weeks apart, are quite low. The authors consider them high enough, however, for group evaluation.

The last paper of this group is concerned with the use of the computer for the assessment of essays. Whalen, starting from previous research at a relatively mature level, selected certain variables and added others to develop a computer program that would be appropriate for seventh-grade essays. Using a sample of seventy-one essays, Whalen shows his computer method for predicting language ability, as measured by the California Language Test, to be quite satisfactory. The California subscores are more difficult to predict accurately. With the rapidly increasing availability of computers, there should be many opportunities to take advantage of such programs as Whalen's and so release teachers for more constructive activity.

Group B. Measurement Specialist. The five papers considered here concern procedures appropriate for large-scale research of a type not ordinarily engaged in by a classroom teacher but, rather, by a full-time specialist. The latter, however, would be well advised not to ignore the papers described under Group A.

Vogt discusses a generalization of the Rasch model to make possible the handling of polychotomously scored items, such as occur in attitude scales. The Rasch model expresses the probability of an individual's response to a test item as a function of the individual's ability and the difficulty of the item. The discussion, however, is brief, concerned with only one aspect--estimation of the parameters--and does not include details of procedure.

Hendrickson and Stanley demonstrate how Guttman's scaling method, which was developed for use with attitude scales where there is no "correct" answer, can be used for differential weighting of options in multiple-choice tests.

Used with a test consisting of two verbal sections and two mathematical sections taken by 10,000 examinees, the method produced scores that yield higher internal consistency reliability estimates than scores obtained by the usual R-KW "correction for guessing" formula. Most, but not all, of the correlations between sections became lower under the differential weighting method. The reader who wishes to use the procedure will find the steps carefully documented.

Gleser discusses the implications underlying the correction of a correlation coefficient for attenuation due to unreliability. In particular, he points out situations where an increase in test reliability may or may not be expected to be accompanied by an increase in validity. It is in the classical single-factor model first introduced by Spearman that the two relations increase together.

Sullins uses simulated data for his investigation of "the effects of sequential dependence (SD) on the sampling distributions of three commonly used reliability estimates--Kuder-Richardson Formula 20 (KR20), Kuder-Richardson Formula 21 (KR21), and odd-even Split-Halves (S-H)." He fails to raise the question of the appropriateness of such formulas when SD occurs. The data permits comparisons of sampling distributions of the three reliability estimates under conditions of sequentially independent items and dependent items for thirty-six combinations of test length, sample size, and distribution of item difficulties. Significant differences among the reliability sampling distributions are found for every comparison presented in the findings.

The versatility of the computer for the research specialist is further illustrated by Lindsay and Prichard, who have developed a Fortran program for the equipercentile method of equating tests. Not only is the method fast and inexpensive but also it is verifiable, since it eliminates the judgment associated with hand-smoothing of a plotted curve.

Group C. Research on Existing Scales or Tests. Grouped here are reports on research that serves to increase knowledge and understanding of existing measurement devices. The first four to be described deal with measurement at levels at or below Grade 3; the next two, with Grades 3 to 6; and the last, at the administrative level.

Olson and Rosen used a method of factor analysis to explore the underlying structure of five reading readiness tests and a reading achievement test. Intercorrelations among thirty-five subscores are not presented. Major factor loadings on correlated factors are listed and discussed.

Data from a behavior rating scale used with migrant children under five years of age was used by Flynn in a factor-analysis study designed to establish the degree of independence and validity of the traits assessed by the scale (Pre-Kindergarten Scale). The factorial analysis of twenty-five items produced four interpretable factors, which were shown to possess a certain degree of validity.

Busse, Blum, and Gutride made use of two forms each of three tests of creativity given to lower-class preschool children in order to study the effects of three different testing conditions. A thoroughly controlled analysis-of-variance procedure led to the conclusion that the creativity measures were largely unaffected by the various testing conditions. Measures presumed to be assessing the same abilities were found to be less highly intercorrelated than their reliability estimates might lead one to expect.

The relations of measures of self-perception-in-school (SPS) with areas of achievement, popularity, and behavior were studied by Alberti, who found a number of small but statistically significant correlations with SPS for girls and boys in Grades 1, 2, and 3. Analyses of variance indicate grade-to-grade and sex differences in SPS. The author stresses the need for further research in this area.

Also dealing with the relation of attitude and achievement is a study of four arithmetic attitude scales by Mastantuono and Anttonen, who show the extent to which an attitude measure (or measures) improves the prediction of achievement over the use of an IQ measure or of teachers' grades alone. The finding of a significant contribution by a measure of arithmetic attitude to the prediction of achievement carries an implication of the responsibility of the classroom teacher to develop favorable attitudes.

A brief report by Bayuk and Proger describes a factor analysis based on the Test Anxiety Scale for Children, administered to about four hundred Grade 6 children. Three to five factors were found among the items of the scale. The paper contains no numerical data.

Boardman describes a computer-based model designed to deal with certain limitations of the "In-Basket" test, a type of test that simulates tasks of administration. Making use of a test designed for the elementary school principal, Boardman gives details of the development of the model and the procedure of feedback to the examinee.

It may be of more than passing interest to note that more than one half of the foregoing papers deal with measurement instruments other than the ubiquitous multiple-choice objective test. Interest in a wide variety of testing devices is a healthy sign. It would seem, however, that modern test development must include some sort of high-speed scoring, an example of which is Whalen's use of the computer for scoring essays. It appears likely, therefore, that either development of new computer or scoring-machine methods may be expected, or measurement scales will evolve into multiple-choice forms.

Papers Reviewed

Alberti, J. M. Correlates of self-perception-in-school. 9p. (ED 048 336, MF and HC available from EDRS).

Bayuk, R. J. Jr., & Proger, B. B. Additional evidence of the multidimensionality of the Test Anxiety Scale for Children. 3p. (ED 046 972, MF and HC available from EDRS).

Boardman, G. R. A computer-based feedback model for an administrative "in-basket" simulation exercise. 12p. (ED 048 368, MF and HC available from EDRS).

Busse, T. V., & Others. Testing conditions and the measurement of creative abilities in lower-class preschool children. 42p. (ED 046 979, MF and HC available from EDRS).

Crocker, L. M., & Mehrens, W. A. The comparative effectiveness of different item analysis techniques in increasing change score reliability. 10p. (ED 049 290, MF and HC available from EDRS).

Flynn, T. M. Convergent-discriminant validation of behavioral ratings. 12p. (ED 049 282, MF and HC available from EDRS).

Gleser, L. J. The attenuation paradox and internal consistency. 16p. (ED 050 142, MF and HC available from EDRS).

Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. 53p. (ED 050 168, MF and HC available from EDRS).

Lindsay, C. A., & Prichard, M. A. An analytical procedure for the equipercentile method of equating tests. 12p. (ED 046 981, MF and HC available from EDRS).

McMorris, R. F. Evidence on the quality of several approximations for commonly used measurement statistics. 19p. (ED 048 373, MF and HC available from EDRS).

Mastantuono, A. K., & Anttonen, R. G. An examination of four arithmetic attitude scales. 22p. (ED 049 294, MF and HC available from EDRS).

Olson, A. V., & Rosen, C. L. Exploration of the structure of selected reading readiness tests. 9p. (ED 044 448, MF and HC available from EDRS).

Peper, J. B., & Chansky, N. M. Development of a scaling technique for sociometric data. 29p. (ED 048 338, MF and HC available from EDRS).

Sullins, W. L. The effect of sequential dependence on the sampling distributions of KR-20, KR-21, and split-halves reliabilities. 9p. (ED 049 276, MF and HC available from EDRS).

Taylor, A. P., & Helmstadter, G. C. A preliminary pair comparison test for measuring aesthetic judgment in young children. 9p. (ED 048 351, MF and HC available from EDRS).

Vogt, D. K. On an extension of the Rasch model to the case of polychotomously scored items. 3p. (ED 049 289, MF and HC available from EDRS).

Whalen, T. E. The analysis of essays by computer: A simulation of teacher ratings. 26p. (ED 048 352, MF and HC available from EDRS).