DOCUMENT RESUME

ED 377 245 TM 022 511

AUTHOR Crehan, Kevin D.; And Others
TITLE Scaling Performance and Objectively Scored Test Items.
PUB DATE Apr 93
NOTE 8p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Curriculum Based Assessment; Educational Testing; Elementary Education; Elementary School Students; Language Arts; Mathematics Achievement; *Objective Tests; Reading Achievement; *Scaling; Scores; *Scoring; *Test Items
IDENTIFIERS *Anchor Tests; *Performance Based Evaluation

ABSTRACT

A strategy is proposed for combining scores from multiple-choice achievement measures with performance assessments. The specific situation discussed involves the revision of a curriculum-based multiple-choice and performance assessment testing program for grades 1 through 6 for a large school district. Reading, language-arts, and mathematics achievement will be assessed at each grade level using multiple-choice tests and at least one performance assessment. Curriculum and assessments were developed locally by teacher task groups. Approximately 48,000 examinees will respond to the multiple-choice tests and complete at least one performance assessment. A scaling test composed of a sample of items from each grade will be constructed and administered to about 300 examinees at each grade level. Multiple-choice tests will be machine scored and about 20% of the performance assessments at each grade level will be scored by trained scorers. The process to develop scaled scores over grade levels and to combine the scores is described in detail. The procedure will attempt to combine common-anchor test design and scaling-test methods. (Contains 10 references.) (SLD)

Scaling Performance and Objectively Scored Test Items

Kevin D. Crehan

University of Nevada, Las Vegas

Thomas M. Haladyna

Robert K. Hess

Arizona State University West

Paper presented at the annual meeting of the *National Council on Measurement in Education*, Atlanta, GA, April, 1993.

Introduction

This is a concept paper which proposes a strategy for combining scores for multiple-choice achievement measures with performance assessments. The specific situation involves the revision of a curriculum based multiple-choice and performance assessment testing program for grades 1 through 6 for a large school district. Reading, language arts, and mathematics achievement will be assessed at each grade level using 30-45 item multiple-choice tests and at least one performance assessment in the areas of reading and mathematics.

The curriculum, multiple-choice items, and performance assessments were all locally developed by teacher task groups with support from curriculum and measurement consultants. The assessment development program has recently finished a large scale field testing for the pools of multiple-choice items and performance assessments and is in the process of developing 'final' multiple-choice test forms, revised performance assessments, and scoring rubrics for May administration and scoring.

The multiple-choice tests are designed to have an overlap of five to seven common-anchor-items between grade levels to allow scaling to achieve comparable scores within content areas. School district administrators wish to have a scaled score in

each content area over the six grade levels. Additionally, they want to combine the multiple-choice results with the performance results for each content area in such a manner as to weight the performance assessment as one-forth the composite. The following presents a proposed strategy to accomplish the scaling of the multiple-choice measures and combine the scaled multiple-choice scores with the performance assessment scores.

Method

There will be approximately forty-eight thousand examinees in grades one through six responding to multiple-choice tests in reading, languate arts, and mathematics. Additionally, examinees will be administered at least one reading performance assessment and one mathematics performance assessment at each grade level. A scaling test composed of a sampling of items from each grade level test will be constructed and administered to approximately 300 examinees at each grade level. The multiple-choice tests will be machine scored and a sample of approximately 20% of the performance assessments at each grade level will be scored by a task group of trained scorers providing at least two ratings per response.

The multiple choice test results will be analyzed using either a polytomous scoring model [e.g., max-alpha (Guttman, 1941) or polyweighting (Sympson, 1983; 1986; 1988)] or an item response

theory IRT model [e.g., three parameter model (Lord & Novick, 1968)]. The BILOG computer program (Mislevy & Bock, 1990) allows test form equating using common anchor or linking items between forms of the test for the 1, 2, and 3 parameter item response theory models. Sympson (1990) developed a program employing a polytomous scoring model which allow equating among test forms with overlapping item sets. Sympson's model is similar to Guttman's (1941) polytomous scoring strategy which provides an optimization of coefficient alpha and, therefore, is often referred to as max-alpha. Max-alpha uses the concept of option mean, the mean of total test score for all examinees choosing an option. Sympson's polyweighting replaces option mean with the mean percentile rank of examinees selecting each option.

Performance assessments will be scored using a scoring model which partitions score variance into facets (Linacre,1989), which allows removal of variability for raters and prompts in the examinees score distribution.

The following presents the plan to develop scaled scores over grade levels and combine performance assessment ratings with the scaled scores. The procedures described will employ cross validation.

Steps in the process:

1. Using either an IRT or classical polytomous model, determine scaled scores over the grade levels for the multiple-choice tests within each content area. The procedure will attempt to combine common-item-anchor-test design (Angoff, 1984) and scaling test methods (Petersen, Kolen, & Hoover, 1989). Mittman (1958) found less grade-to-grade overlap for the scaling test method than for the anchor item method. Both methods will be used and combined to maximize the likelihood of developing an educationally relevant score scale.

2. Next, convert each grade level's distribution of scaled scores on the multiple-choice tests to percentile ranks.

3. Determine the examinee ability logits for the performance assessments using Facets.

4. Determine the percentile ranks of the performance assessment logits within grade levels.

5. Equate performance assessments logits to multiple-choice scale scores using equipercentile equating.

6. Score each examinee by determining the linear combination of the multiple-choice scale score and performance assessment score using appropriate weighting.

7. Convert the combined score to a score scale with desirable characteristics.

8. Convert combined scaled scores to percentile ranks within grades.

Conclusion

The primary focus of the curriculum and test revision project is to improve instruction and learning and the primary interpretations of test results will be curriculum (or criterion) based. The scaled scores, therefore, will not be an initial focus of test interpretation. It is most likely that, if the scaling is judged to be somewhat successful, an effort to improve the tests and scaled scores will follow.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, N. J.:Educational Testing Service.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.) *Prediction of Personal Adjustment. Social Science Research Bulletin, 48,* 321-345.

Linacre, J. M. (1989). *Many Faceted Rasch Measurement.* Chicago, IL: MESA Press.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3 Item analysis and Test Scoring with Binary Logistic Models (2nd ed.).* Mooresville, IN: Scientific Software.

Mittman, A. (1958). An empirical study of methods of scaling achievement tests at the elementary grade level. Unpublished doctoral dissertation, The University of Iowa, Iowa City.

Petersen, N. S., Kolen, M. J., and Hoover, H. D. (1989). Scaling, Norming, and Equating. In R. L. Linn (Ed.), *Educational Measurement, (3rd ed., pp. 221-262). New York: American Council on Education-Macmillan.*

Sympson, J. B. (1983). A new item response theory model for calibrating multiple-choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.

Sympson, J. B. (1986). Extracting information from wrong answers in computerized adaptive testing. Paper presented in C. E. Davis (Chair), New Developments in Polychotomous Scoring. Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Sympson, J. B. (1989). A procedure for linear polychotomous scoring of test items. Paper presented at the 1988 Office Naval Research Contractors' Meeting on Model-Based Psychological Measurement, Iowa City, IA, May 1988.

Sympson, J. B. (1990). POLY: A computer program for polychotomous item analysis. San Diego, CA: Navy Personnel Research and Development Center.