DOCUMENT RESUME

ED 284 911 TM 870 496

AUTHOR Hambleton, Ronald K.
TITLE Evaluating Criterion-referenced Tests. ERIC Digest.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 86
CONTRACT 400-86-0018
NOTE 4p.
AVAILABLE FROM ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, NJ 08541-0001 (single copy free).
PUB TYPE Information Analyses - ERIC Information Analysis Products (071) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Behavioral Objectives; *Criterion Referenced Tests; Decision Making; *Evaluation Criteria; *Specifications; *Test Construction; Test Interpretation
IDENTIFIERS *ERIC Digests; Standards for Educational and Psychological Tests; Test Specifications

ABSTRACT

Criterion-referenced tests (CRTs) are constructed to permit the interpretation of examinee tests performance in relation to a set of well-defined competencies. CRTs are currently used extensively in schools, industry, and the armed services because they provide valuable and different information from norm-referenced tests. Test publishers, school districts, and state departments of education produce CRTs; however, many of the available tests fall far short of the technical quality necessary for them to accomplish their intended purposes. This digest provides practitioners and test developers with guidelines for evaluating CRTs. Drawn from the Standards for Educational and Psychological Testing, 25 content and technical questions are presented that must be answered when evaluating criterion-referenced tests. The technology for preparing CRTs is now well developed, and practitioners can avoid improperly prepared tests by addressing these questions. (BS)

ERIC DIGEST

Evaluating Criterion-referenced Tests

ERIC Clearinghouse on Tests, Measurement, and Evaluation

Educational Testing Service

Princeton, NJ 08541-0001

(690) 734-5181

ERIC DIGEST

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service, Princeton, NJ 08541-0001 (609) 734-5181

Evaluating Criterion-referenced Tests

Criterion-referenced tests (CRTs) are constructed to permit the interpretation of examinee test performance in relation to a set of well-defined competencies (Popham, 1978). CRT scores have three common uses:

1. to describe examinee performance in relation to competencies of interest;
2. to assign examinees to mastery states (e.g., "masters" and "nonmasters"), for each competency of interest, or in relation to a group of competencies defining a domain of content; and
3. to describe the performance of specified groups of examinees in program evaluation studies.

CRTs are currently used extensively in schools, industry, and the armed services because they provide valuable information that differs from the information provided by norm-referenced tests (NRTs). But CRTs, like other data-collection instruments used in educational decision-making, are of variable quality, and lesser quality tests are not going to fully meet the informational needs of users. This digest was prepared to help practitioners identify high quality criterion-referenced tests. Of course the same guidelines should be useful to test developers as well.

BACKGROUND

Most of the major test publishers have available an assortment of criterion-referenced tests for assessing reading, mathematics, language arts, and other content areas in grades K to 12. In addition, many local school districts, state departments of education, and smaller test publishers have produced their own criterion-referenced tests. Many of the available tests, however, fall far short of the technical quality necessary for them to accomplish their intended purposes. When tests lack sufficient technical quality, there are a number of plausible explanations: For one, many of the available criterion-referenced tests were developed before an adequate testing technology was fully explicated. Fortunately, an adequate technology for constructing criterion-referenced tests and using criterion-referenced test scores is now available (Berk, 1984; Hambleton, in press; Hambleton, Swaminathan, Algina, & Coulson, 1978; Popham, 1978). Guidelines can be produced by which criterion-referenced tests and their manuals can be evaluated. The recently published *Standards for Educational and Psychological Testing* (1985) for evaluating tests and test manuals, prepared by a joint committee of AERA, APA, and NCME, is helpful, too, and was used in preparing the next section.

TEST EVALUATION

There are 25 content and technical questions that must be answered when evaluating criterion-referenced tests, commercially prepared or otherwise:

Content Questions

1. Do the competencies measured by the test cover the content domain of interest?
2. Are the competencies themselves well-defined so that the appropriate domain of content for each competency is clear?
3. Is there a capability of adding to or taking away from the test content so that the final test provides a suitable match to the content domain of interest?
4. Is an appropriate rationale offered for the selection of competencies measured in the test?
5. Is the test-item content appropriate to measure the competencies?

Technical Questions

6. Do the test items meet the standard item-writing principles?
7. Are the test items free from bias and stereotyping?
8. Is each group of test items measuring a competency *representative* of the domain of content spanned by the competency?
9. Was the item-review process carried out properly?
10. Was a suitable sample of examinees used to pilot the test items?
11. Were item statistics used correctly in building the test?
12. Do the test directions address important information such as test purpose, scoring, time

limits, passing score(s), and marking answer sheets (or test booklets)?

13. Are the time limits sufficient for examinees to complete the test?
14. Are the test administrator's directions complete so as to insure a proper test administration?
15. Are the print size, quality of printing and artwork, and page layouts appropriate for the examinees?
16. Are the reliability and validity studies conducted with large enough samples of examinees for whom the test is intended?
17. Are useful reliability indices, such as "decision-consistency" and 'kappa," reported for the test scores?
18. Are the reliability indices high enough to justify the use of the test in the intended application?
19. Are personal and environmental factors that influence test performance addressed in the test manual?
20. Is a test manual available that addresses test purposes, development, administration, scoring, psychometric properties of the test scores, and test interpretations?
21. Is there justification offered (and is it appropriate) for the choice of standard (or cut-off score)?
22. Is the process used to set a standard fully documented in the manual, and is it appropriate?
23. Is there acceptable and fully documented validity evidence for the intended use(s) of the test scores?
24. Are there cautions in the technical manual about the size of errors of measurement and/or misclassification and the role of these errors in score interpretations?
25. Are the test scores reported fully and clearly?

Clarification and expansion of many of the questions above can be found in Berk (1984), Hambleton (in press), and Popham (1978).

CONCLUDING REMARKS

Identifying well-constructed, reliable, and valid criterion-referenced tests is essential for insuring that the purposes of a testing program are accomplished. The importance of the 25 individual questions above will vary somewhat from one test to another. Still, some attention to each question in criterion-referenced test evaluation would normally be desirable. The technology for preparing criterion-referenced tests is well-developed at this time. Practitioners should expect that the technology will be used and used correctly in preparing tests, and when it is not, these improperly prepared tests should be avoided.

Ronald K. Hambleton,
University of Massachusetts
1986

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological tests.* Washington, DC: APA.

Berk, R. J. (Ed) (1984). *A guide to criterion-referenced test construction*. Baltimore, MD: Johns Hopkins Press.

Hambleton, R. K. (in press). *Criterion-referenced testing methods*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D. B. (1978). Criterion-referenced testing and measurement: Review of technical issues and developments. *Review of Educational Research*, 48, 1-47.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

OERI
Office of Educational Research and Improvement
U.S. Department of Education

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education under contract no. 400-86-0018. The opinions expressed in this report do not necessarily reflect the position or policies of OERI or The Department of Education.

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service, Princeton, NJ 08541-0001 (609) 734-5181

7290502 • Y67P.5 • Printed in U.S