DOCUMENT RESUME

ED 480 064 CG 032 637

AUTHOR Camara, Wayne J.

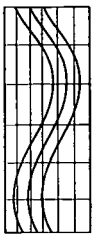
TITLE Professional Testing Standards: What Educators Need To Know.
PUB DATE 2003-08-00
NOTE 11p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE Information Analyses (070)
EDRS PRICE EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Testing; *Psychological Testing; Scoring; *Standards; *Test Use; *Testing Problems

ABSTRACT

Real and perceived misuses of educational tests, errors in test scoring and test use, and incidents of cheating on tests have been widely reported in local and national media. As educational tests take on additional importance for students, teachers, and schools, there is appropriate concern about the quality of assessments and the appropriate use of tests and test data. Given this situation, testing standards that represent professionals in educational measurement and psychology have increasing importance in evaluating test use today. The American Educational Research Association, the American Psychological Association, and the National Council for Measurement in Education completed their fourth collaboration in producing the "Standards for Educational and Psychological Testing" in 1999. This chapter provides an overview of the issues addressed in the current standards and their relevance to educators, as well as briefly describes the development of these standards and how they may be used today. (GCP)

Professional Testing Standards: What Educators Need to Know

By
Wayne J. Camara



Chapter 29
Professional Testing Standards
What Educators Need to Know

Wayne J. Camara

Children are often first exposed to standardized testing in elementary school, and by age 18 assessments may have played a major role in their life decisions, ranging from graduation and promotion to college admission and entry into certain majors or selection into occupations or organizations. Real and perceived misuses of educational tests, errors in test scoring and test use, and incidents of cheating on tests have been widely reported in local and national media (Camara, 1997). As educational tests take on additional importance for students, teachers, and schools, there is appropriate concern about the quality of assessments and the appropriate use of tests and test data.

At times misuse of tests has resulted in legal challenges to state, district, or school assessment practices. In some instances, concerns about testing practices have also resulted in legislation, such as test disclosure laws requiring the release of some test forms in some states. In addition, federal legislation has often included language concerning the types of assessments used and their role in relation to federal funding of educational initiatives. Other federal laws strive to protect certain groups from specific abuses (e.g., the Americans With Disabilities Act of 1990, the Civil Rights Act of 1991). However, the majority of concerns regarding the quality of tests and the appropriate use of tests in education are matters of professional practice and technical, or psychometric, concern. Given this situation, testing standards that represent professionals in educational measurement and psychology have increasing importance in evaluating test use today.

Background

The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council for Measurement in Education (NCME) completed their fourth collaboration in producing the *Standards for Educational and*

Psychological Testing in 1999. This chapter will provide an overview of the issues addressed in the current standards and their relevance to educators, as well as briefly describe the development of these standards and how they may be used today.

In 1954, APA issued the first set of testing standards, entitled *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Parallel standards were developed for educational achievement tests in 1955 by the American Educational Research Association and the National Council for Measurement Used in Education (later renamed the National Council for Measurement in Education; Camara and Kraiger, 1996). AERA, APA, and NCME collaborated on joint standards that incorporated educational and psychological testing in 1966; they issued revisions in 1974 and 1985, and completed the current *Standards for Educational and Psychological Testing* in 1999.

The *Standards for Educational and Psychological Testing* (referred to as the *Standards*) is the most widely cited document addressing technical, policy, and operational issues for educational assessment (Camara, 1997). Yet most policymakers and educators with responsibilities for assessment practices may not be familiar with the *Standards* because they are not members of these three associations and because the document is primarily technical in nature and is not likely to be used in introductory testing or measurement courses or workshops that these audiences may frequent. In an effort to address this concern, the Joint Committee on Testing Practices developed the *Code of Fair Testing Practices in Education* (2002). This code attempts to highlight key concepts from the nearly 200-page *Standards* in a four-page brochure that professionals are encouraged to disseminate.

The *Standards* is based on the premise that effective testing and assessment require that all participants in the testing process possess the knowledge, skills, and abilities relevant to their role, as well as awareness of personal and contextual factors that may influence the testing process. "Although the evaluation of the appropriateness of a test or testing application should depend heavily on professional judgment, the *Standards* provides a frame of reference to assure that relevant issues have been addressed" (AERA et al., 1999, p. 2). The term *test* refers to a broad range of instruments and measures, and the standards apply regardless of the specific label applied to the instrument (e.g., assessment, scale, inventory). The only distinction made regards standardization. The authors acknowledge that the *Standards* applies primarily to any standardized measure and only to a lesser degree to

nonstandardized methods (e.g., unstructured behavior samples, teacher-made tests).

The *Standards* document contains 15 chapters and 264 standards divided into three major sections: (a) Test Construction, Evaluation, and Documentation, (b) Fairness in Testing, and (c) Testing Applications. The remainder of this chapter describes the major concepts addressed in the three sections and their implications for educators.

Test Construction, Evaluation, and Documentation

This section focuses primarily on the responsibilities of test developers and test users and addresses psychometric issues such as validity, reliability, test development, norms, test administration, scoring, and documentation. The *Standards* defines validity as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (AERA et al., 1999, p. 184). Validity is the most important consideration in developing and evaluating educational assessments. Validation involves accumulating the evidence that provides a scientific basis for the proposed test score interpretations. It is the interpretations of these scores that are evaluated, not the assessment itself. For example, if an achievement test is developed and used for student placement into advanced math courses, evidence supporting the validity of such use is required. One source of evidence may be a finding that performance on the achievement test is related to performance in the subsequent mathematics courses. If a cutoff score or specific performance level is used to make the placement decision, one would want additional evidence that students who perform below the cutoff point are less likely to succeed in the advanced courses than students who perform above that point. When evidence is not adequately compelling, additional measures such as teacher recommendations, student or parent recommendations, and academic performance should be included. In another instance an achievement test may be used to compare the writing skills of eighth graders in a state over a number of years to determine if standards-based reform activities are improving student performance. Because the achievement test is used to make inferences about the comparability of scores for groups of test takers, rather than individual students, different sources of evidence are required. Although more than one source of validity evidence is generally desirable, the quality of evidence is of primary importance. The intended uses of the assessment and the proposed interpretation of scores have implications

for test development and evaluation.

The *Standards* notes that a test itself is not validated, but rather the intended use of the test and how test scores will be interpreted are validated. Five sources of evidence that can contribute to a validation strategy are listed: (a) content, (b) relationships between test scores and other variables, such as test-criterion relationships, (c) internal structure of the test, (d) response processes, and (e) consequences of testing.

Test content. Content includes the test items or performance tasks, format and wording of questions, response formats, and instructions for administration and scoring. Evidence based on test content should demonstrate that the test content is aligned to the curriculum taught or the skills required for future success (e.g., placement).

Relationships between test scores and other variables. A typical study examines how accurately test scores predict criterion data at a later time (e.g., admissions testing predicting performance on college coursework), whereas a concurrent study collects predictor and criterion data during a relatively short time frame. Such studies may be used to determine (a) if the relationship between the predictor (e.g., test, grades) and the criterion (or outcome measure; e.g., freshman grades, graduation) differs across subgroups, or (b) the accuracy of a test for admission or placement decisions.

Internal structure of the test. Such evidence examines any relationships among test items or tasks that can provide additional evidence of how test scores may relate to specific aspects of the construct that is to be measured.

Response processes. Evidence based on response processes may be collected by examining the processes that test takers use in responding to test questions or tasks. Often analyses of individual responses can be gathered by questioning test takers about their strategies in responding to a specific question, through examinee responses on computerized assessments, or through experimental studies.

Consequences of testing. Although evidence regarding consequences may influence decisions concerning the use of an assessment or other measure, it will not usually be related to inferences concerning the validity of scores. For example, group differences in performance on

an assessment are relevant to a school or institution, yet such differences alone do not necessarily detract from the validity of intended test interpretations.

One of the most important issues for educators in the *Standards* is the discussion about the respective responsibilities of the test developer and test user in accumulating validation evidence relevant to the intended use of the test and the inferences that will be made from test scores. Specifically, the *Standards* states that the test developer should clearly set forth the intended uses for a test, but if a test user wishes to use the test in a way for which sufficient evidence has not been presented, then it is that user's responsibility to provide such evidence. For example, if a test is developed to provide diagnostic information about aggregate groups of students. but a user (e.g., school district, state) decides to use the test to determine student promotion, then it is incumbent on the user to provide evidence to support that new use. If the test is used to classify students into proficiency groups or to assign students to different educational programs or courses, validation evidence for such classifications is required. That is, it is not adequate simply to demonstrate that there is a relationship between the test and some criterion (e.g., grades); rather, evidence supporting the validity of the classification decision is needed. Similarly, the *Standards* notes that when score differences are used to distinguish groups, such as students classified as proficient versus exemplary in an area, the reliability of the data, including the standard errors or confidence intervals for scores, should be reported along with the test score.

This section of the *Standards* also elaborates on procedures for administration, scoring, and interpretation of tests. It addresses issues such as retention of student test scores, errors in testing materials, disruptions in standardized administrations, procedures for challenging test scores, human raters or scores, and the types of documentation that should be provided in a testing program. Finally, a discussion of score reliability and test development issues concerning performance assessments, portfolios, and other educational assessments is provided.

Fairness in Testing

The Fairness in Testing section addresses issues of fairness and bias in testing and includes separate discussions of test takers' rights and responsibilities, the testing of individuals with diverse language backgrounds, and the testing of individuals with disabilities. The *Standards* discusses four different aspects of fairness. The first two

relate to ensuring that tests are absent any bias and to the need to treat all examinees in an equitable fashion in the testing process. The third component of fairness is that all subgroups (e.g., based on ethnicity, race, gender, disability) must have equal passing rates or scores. The *Standards* acknowledges that there is broad consensus on the first two aspects, but that the idea that equal outcomes among groups is required for fairness "has been almost entirely repudiated in the professional testing literature" (p. 74). The fourth component of fairness concerns opportunity to learn.

The *Standards* provides a detailed discussion of how item bias and predictive bias could represent major challenges to the technical qualities of a test, and they describe procedures to ensure equal treatment of all students in testing. If some students have not had the opportunity to learn what is assessed in an achievement test, the scores reflect what the test taker knows but also what he or she has not had an opportunity to learn. When students have not had an opportunity to learn all tested information, then any policy about, for example, using the test scores as a basis for withholding a diploma is unfair.

The *Standards* also discusses some of the threats to the validity of inferences made from test scores of students who may not be proficient in the language in which they were tested. The greatest threat may occur when students' language proficiency limits their performance in an area other than language proficiency. Many state tests employ extended reading passages and written responses to demonstrate proficiency in areas such as mathematics, science, or history. To the extent that such assessments rely on language skills, students' scores may not accurately reflect their knowledge in these areas, but may instead reflect a combination of knowledge gaps and poor language proficiencies. The *Standards* describes four types of modifications that are designed to accommodate students with disabilities: (a) presentation format (e.g., large print, cassette), (b) response format (e.g., computer keyboard, aide to record oral responses), (c) extended time, and (d) test setting (e.g., individual administrations). Test users should take steps to ensure that test scores, and the inferences made from them, reflect the intended construct rather than the disabling condition. For example, if a student who needs longer than average time for cognitive processing is required to complete a speeded test, the results may in part reflect the disability. On the other hand, any modifications made should be described in detail and, when feasible, evidence should be provided that the inferences drawn from the results are valid and comparable to inferences based on scores of students who did not receive the

accommodation. When testing students with disabilities or limited English proficiency, it is often difficult to demonstrate comparability of scores between those students and other test takers, and professional judgment is required in making inferences from these scores.

Testing Applications

The final section of the *Standards* describes the responsibilities of test users followed by the application of testing in specific settings. Of most relevance to educators are chapters discussing educational testing and assessment and the role of testing in program evaluation and public policy. In this section, a number of important points made earlier in the document are described more fully as they apply to education and public policy:

- Many tests are used for multiple purposes; however, evidence needed to support one use (e.g., program goals) will differ from evidence required for another purpose (e.g., individual student use).
- The higher the stakes associated with a test, the more important it is that test-based inferences be supported with strong technical evidence.
- Performance assessments often require complex procedures and training to increase the accuracy of scorers' judgments, and coverage of content domains is often reduced because each task usually requires more time to complete than do objectively scored items.
- When a test is intended to serve as an indicator of student achievement of curriculum standards, evidence of the extent to which the test samples the range of standards is needed.
- A decision that will have major impact on a student should not be made on the basis of a single test score; other relevant information should be considered in conjunction with the score.
- Individuals who supervise testing should have the necessary education and training to ensure they are familiar with the evidence for the validity and reliability of the test for the uses they intend.
- When schools, districts, states, or other authorities mandate the use of certain tests, those entities are responsible for identifying and monitoring the impact of testing and to minimize potential negative consequences.

- The integrity of test results must be maintained by eliminating practices that could raise scores without improving performance.

Summary

The *Standards* represents an important resource for all educators and policymakers who use and interpret test scores for individual students or groups of students (e.g., in a school, district, state, or nation). First, it represents the consensus of professionals in psychological and educational testing. The standards were developed by a committee composed exclusively of academicians and other researchers who have expertise in testing, measurement, and education, and their purpose is to provide guidance to all professionals who develop, select, or use assessments. Second, it is designed to promote sound and ethical use of tests by providing rigorous standards, some of which may not be feasibly met in many settings, that assist educators in evaluating test quality and appropriate test use. Third, it is based on current scientific knowledge and professional practice. Finally, it provides detailed discussions of several possible conflicts and concerns, ranging from issues that are highly technical and psychometric to those that concern proper administrative procedures or documentation and communications about a testing program. The *Standards* contains an extensive list of organizations and individuals with expertise in testing and education who reviewed, contributed to, and in many instances have endorsed these standards. Given the increased use of educational tests and the role they play in the allocation of resources and accountability in education, it is vital that educators and policymakers concerned with these issues become familiar with the professional and technical requirements related to testing in order to reduce misuse of tests and test results.

References

AERA, APA, & NCME. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: APA.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Educational Research Association & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests.* Washington, DC: Authors.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

Camara, W. J. (1997). Use and consequences of assessments in the U.S.A.: Professional, ethical, and legal issues. *European Journal of Psychological Assessment, 13*(2), 140–152.

Camara, W. J., & Kraiger, K. (1996). Organisational infrastructure for selection and assessment in the U.S.A. In M. Smith & V. Sutherland (Eds.), *International review of professional issues in selection and assessment* (Vol. 2, pp. 138–146). Chichester, UK: John Wiley & Sons.

✦Joint Committee on Testing Practices. (2002). *Code of fair testing practices in education.* Available on *Measuring Up: An Anthology of Assessment Resources* [CD]. Also retrievable on-line: http://aac.ncat.edu.

✦ Document is included in the Anthology of Assessment Resources CD

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").