

Standards for Educational and Psychological Testing:  
Influence in Assessment Development and Use

Wayne J. Camara

The College Board

Unpublished Paper

November 8, 2007

The appropriate development and use of assessments are essential requirements for responsible professional practice in educational testing and measurement. The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have collaborated on the joint development of *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*) since 1966. There have been four revisions to these joint standards since they were first issued as separate technical recommendations for achievement tests and psychological tests by AERA and NCME in 1955, and APA in 1954. The most recent edition of the *Standards* was published in 1999, and another revision is scheduled to begin in 2008.

When the *Standards* were developed the test user was considered to be a trained professional who generally had some graduate training and supervised experience in assessment. These “primary test users” would include test developers, testing contractors, state and district assessment directors, and school counselors who develop or use tests for decision making purposes (Camara, 1997).

Today, the term test user encompasses a much broader group of “secondary test users,” including individuals with little or no training in measurement and assessment such as teachers, parents, policymakers, and the media (Camara, 1997). Policymakers and educational administrators may have great influence over the use of assessment results and may misuse assessments in today’s accountability environment (Berliner & Biddle, 1995). The further the test users are from the assessment, the less familiar they may be with the intended use of the assessment, evidence supporting the validity of inferences concerning the use of assessment results, and test content and characteristics of the test taking population, which increases the likelihood that test misuse will occur (Camara and Lane, 2006).

#### Purpose and Use of the *Standards*

The *Standards* have continued to emphasize that their primary purpose is to provide criteria for evaluating tests and testing practices and to encourage test developers, sponsors, publishers, and users to adopt the *Standards*, but there is no requirement on members of the professional associations or testing organizations and users to do so. They also note that the *Standards* do not attempt to provide psychometric answers to policy or legal questions. In 1999, the *Standards* abandoned the former designations of each standard as *primary* (required for all tests before operational use) *secondary* (desirable, but not required) or *conditional* (applicable in some instances and settings) (AERA et al., 1999). This change met with some criticism and controversy because it appeared to remove any absolute criteria or requirements for testing and test use and relied more on professional judgment in adherence to each standard.

The *Standards* also apply broadly to a wide range of standardized instruments and procedures that sample an individual's behavior that can include tests, assessments, inventories, scales, etc. The main exceptions in applying the *Standards* are for unstandardized questionnaires (e.g., unstructured behavioral checklists or observational forms), teacher-made tests and subjective decision processes (e.g., teacher evaluating classroom participation over the semester). The *Standards* apply equally to standardized multiple-choice tests as they do to performance assessments (including tests comprised only of open-ended essays) and hands-on assessments or simulations.

There is no mechanism to enforce compliance to the *Standards* on the part of the test developer or test user. Today, many tests are sold and marketed that do not provide documentation required concerning their appropriate use, validation evidence to support such uses, and basic technical documentation such as the reliability of the score scale or a description of the normative or standard setting samples used for score reporting. Some publishers have ignored requests for technical manuals or validation studies citing the proprietary nature of their clients while some test users have used tests for unintended and multiple purposes with no concern for collecting additional evidence to support such uses. Requests for proposals from states and local educational departments nearly always reference the *Standards* and frequently include a broad statement that ‘vendors responding to the RFP must comply’ with the *Standards*; yet few states have conducted detailed audits of their assessment programs in direct reference to all the applicable standards. Wise (2006) describes how technical advisory committees (TACs) and the peer review process used by the U.S. Department of Education for assessment systems under No Child Left Behind (NCLB) are efforts to improve the quality of testing, but do

not base reviews on all relevant components of the *Standards*. Madaus, Lynch and Lynch (2001) and Kortz (2006) have described the need for some independent mechanisms to interpret, encourage compliance with, or even enforce the *Standards*. However, the *Standards* have been referenced in law and cited in Supreme Court and other judicial decisions lending additional authority to the document. For example, they have been cited in Goals 2000: Educate America Act<sup>1</sup> and Title I (Elementary and Secondary Education Act)<sup>2</sup>. They were also cited in several major court decisions involving employment testing, including a Supreme Court case in 1988<sup>3</sup>.

#### Additional Standards and Guidelines in Testing

The American Psychological Association (APA) adopted the first formal ethics code for any profession using assessments in 1952. Eighteen of approximately 100

---

<sup>1</sup> PL 103-227 – Goals 2000: Educate America Act – Sec. 211 Purpose States “the National Education Standards and Improvement Council shall. (5) certify State assessments submitted by States or groups of States on a voluntary basis, if such assessments – (A) are aligned with and support State content standards certified by such Council; and (B) are valid, reliable, and consistent with relevant, nationally recognized, professional and technical standards for assessment when used for their intended purposes.” The Federal Register, 43, pp. 38290-38315.) defines assessment under Goals 2000 act as, “ASSESSMENT – Any method used to measure characteristics of people, programs, or objects. (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. [1985]. *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.)”

<sup>2</sup> Title I – ESEA (Guidance on Standards, Assessments, and Accountability) “Title I requires that the assessment system be used for purposes that are valid and reliable and that it meet nationally recognized professional and technical standards...The primary reference for technical quality of educational assessments is *Standards for Educational and Psychological Testing* (1985), developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.”  
(see [http://www.ed.gov/policy/elsec/guid/standardsassessment/guidance\\_pg4.html](http://www.ed.gov/policy/elsec/guid/standardsassessment/guidance_pg4.html)).

<sup>3</sup> WATSON v. FORT WORTH BANK & TRUST, 487 U.S. 977 (1988) 487 U.S. 977

principles in that Code (APA, 1953) addressed issues such as qualifications of test users, security of testing materials, documentation required in test manuals, and responsibilities of test publishers and test users. Ethical standards for assessment are one of nine areas addressed by the current code (APA, 2002). Many other professional associations with members involved in assessment have similarly adopted ethical standards and professional codes in the past two decades. Increased public awareness of ethical issues, the variety of proposed and actual use of assessments, and the increased visibility and importance placed on assessments for accountability have resulted in greater attention to ethical and professional responsibilities by many professional associations (Eyde & Quaintance, 1988; Schmeiser, 1992).

In the early 1990s the American Counseling Association (ACA) and the American Educational Research Association (AERA) each approved ethical standards that cover a broad range of standards for behavior in counseling and educational research, but make only passing reference to assessment. Ethical standards of ACA (1998), APA, and the National Association of School Psychologists (NASP, 1997; 2000) are unique in that these associations support formal enforcement mechanisms that can result in suspension and expulsion of members, respectively (Camara, 1997)<sup>4</sup>. Ethical standards were first adopted by AERA in 1992 and twice revised. The current standards (AERA, 1999) are designed to guide the work of educational researchers but are not enforceable.

In contrast to laws and regulations that are designed to protect the public from specific abuses, ethical standards and codes attempt to establish a higher normative

---

<sup>4</sup> The National Association for College Admission Counseling (NACAC) has developed standards of practice and other policy guidelines that are enforceable to its institutional members.

standard for a broad range of professional activities and behaviors. For example, APA's Ethics Principles state "if this Ethics Code establishes a higher standard of conduct than is required by law, psychologists must meet the higher ethical standards (2002, p. 1062)." ACA, AERA, APA and the Society for Industrial and Organizational Psychology (SIOP) have followed up the development of ethics codes with casebooks that attempt to guide users in interpreting and applying their standards.

The increased use of tests for accountability has also increased the urgency of informing and educating secondary users of their responsibilities in the appropriate and ethical use of tests and test data. In 2000, the U.S. Department of Education's Office of Civil Rights drafted a resource Guide on High Stakes Testing for educators and policy makers (2000) that attempted to interpret the technical and professional testing standards and legal principles and apply them to high stakes uses in schools, but the guide has not been disseminated since a change of administrations. *Standards for Educational Accountability Systems* (CRESST, 2002), which attempt to apply professional standards to accountability systems for a broad group of educators, have been developed after passage of educational reform law.

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988; In press) attempts to condense the most salient statements concerning the responsibilities of test users and test developers from existing codes and standards in four areas: (a) development and selection of tests; (b) administration and scoring of tests; (c) reporting and interpretation of test results; and (d) informing test takers. The Code has been endorsed by most of the major test publishers and is frequently reproduced on Web pages and publications in an attempt to guide educational professionals in appropriate

practice and use of assessments. Most other technical and professional standards have a much more limited distribution, primarily to members, while the Code encourages reproduction and dissemination. A similar document, *Responsibilities of Users of Standardized Tests* (Association for Assessment in Counseling, 2003), was developed to enhance ethical standards and assist counselors in the ethical practice of testing.

### Test Development

Ethical and professional issues relating to the development of assessments is a broad area that includes test construction (the technical qualities of assessments, evidence supporting the validity of inferences made from test results, and norms and scales) as well as modifications to the test, technical documentation, statements and claims made about assessments, and appropriate use of copyright materials. Technical competence in the development and selection of assessments is an ethical issue. APA's Ethics Code states that persons who develop tests "use appropriate psychometric procedures and current scientific or professional knowledge for test design, standardization, validation, reduction or elimination of bias and recommendations for use" (2002, p. 1072). Test developers are responsible for ensuring that assessment products and services meet applicable professional and technical standards and should be familiar with the *Standards* and other applicable requirements. They also have a responsibility for providing technical documentation on their tests, including evidence of reliability and validity that supports inferences that will be made from test scores. Technical qualities for many educational tests also include construct representation and curriculum relevance (Messick, 1989). Educators who select among "off-the-shelf" tests should ensure that the content specifications of the tests are aligned to the curriculum and that assessment

formats are relevant. The *Standards* reiterate the importance of construct relevance as a central requirement in the validation argument (AERA, et al., 1999).

Test developers need to employ appropriate processes for item development, review, and test assembly to ensure potentially offensive (or biased) content or language is avoided and test content is relevant for the intended use. Evidence that differences in performance across major subgroups are related to the construct being measured and not due to construct irrelevant variance is also a professional responsibility of developers (Joint Committee on Testing Practices, 2004). Test development also includes obtaining appropriate permissions when copyrighted texts or art work are used for an assessment. The *Standards* reiterate that tests should be developed on a sound scientific basis (AERA, et al., 1999).

There are a number of potential dilemmas that may arise for test developers who produce “off-the-shelf”, as well as those who purchase these tests. For example, a test contract may require a developer to produce many more items in a short period of time than the organization is capable of developing while meeting acceptable quality standards. The increased assessment demands from local, state, and federal arenas may require test developers to take actions to meet scheduling and economic constraints that can threaten the technical quality of an assessment program. Sometimes the demand of test production may outstrip the resources of a test publisher and result in errors that may have been prevented with a more reasonable schedule (Phillips and Camara, 2006).

Test users who attempt to use a test for multiple purposes must provide evidence to support the use of the test for each proposed purpose unless existing evidence has been provided by the publisher or other sources. Educators often propose using the same test

for both formative and summative purposes or to provide student, teacher, and school accountability functions as well as instructional, diagnostic, and placement purposes for the student. Evidence to support each specific use should be provided according to the *Standards*.

Rhoades and Madaus (2003) report on several instances where insufficient piloting and pretesting led to spurious results, and time schedules for accountability tests didn't allow for all the quality control procedures needed to detect and correct errors prior to test administration. In some situations, the desire to have tests drive the curriculum and may conflict with any accountability uses for test results. For example, math teachers recruited to determine test specifications for a state accountability test may insist that a new 8<sup>th</sup> grade assessment require the use of graphing calculators to “send a signal” to students and all teachers about the need to introduce students to this technology and to compel the state to support the purchase of graphing calculators and appropriate professional development. However, if students are required to take the test before they are proficient in the use of graphing calculators, then the assessment is measuring construct irrelevant variance. Opportunity to learn is an important component of tests designed to measure achievement. In each of these instances, the technical quality of the assessment raises professional and even ethical issues that are more serious when the stakes associated with the testing program are higher.

The development of standardized administrative procedures and appropriately modified forms of tests and administrative procedures is also a responsibility of test developers (AERA et al., 1999; Joint Committee on Testing Practices, 2006; NCME, 1985). Modifications in test forms, response format, and test setting or content for

students with disabling conditions or diverse linguistic backgrounds should be clearly described for test users to reduce potential misuse of assessments.

Finally, technical documentation is a major requirement of any testing program. The first technical recommendations and standards produced by a professional association (APA, 1954) were largely developed to address the concern that tests were often released without adequate supporting documentation and research (Novick, 1981). *NCME's Code* (1985) notes that current technical information to support the reliability, validity, scoring and reporting processes, and other relevant characteristics of the assessment should be made available to the appropriate persons. The *Code of Fair Testing Practices in Education* (2004) and *Standards* (AERA et al., 1999) add that technical information provided to users should also include the level of precision of test scores, descriptions of test content and skills assessed, and representative samples of test items, directions, answer sheets, and score reports.

### Conclusion

The increased importance placed on the use and results of high stakes assessments has not only placed enormous pressure on students, teachers, principals, school boards, and other parties inside the educational system, but has created additional professional and ethical demands on measurement and testing professionals. The increased stakes associated with educational testing has led to a variety of unintended consequences that impact curriculum, teacher training, and professionalism in areas not directly tied to assessment as well as errors in the testing process itself (Cizek, 2001; Phillips and Camara, 2006). The *Standards* (AERA, et al., 1999) and other professional, technical, and ethical guidelines that address assessment professionals provide guidance on many of

the issues measurement professionals have long encountered in test development, test use, and test reporting.

References:

American Educational Research Association (2000). Ethical standards of AERA.

Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: AERA.

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. American Psychologist, *57* (12), pp. 1060-1073.

American Psychological Association (1953). Ethical standards for psychologists.

Washington, DC: Author.

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

Association for Assessment in Counseling (2003). Responsibilities of Users of Standard Tests (RUST). Alexandria, VA: Author.

Berliner, D. C. & Biddle, B. J. (1995). The manufactured crisis: Myths, fraud and the attack on America's public schools. Reading, MA: Addison-Wesley.

Camara, W. J. (1997). Use and consequences of assessments in the USA: Professional, educational and legal issues. European Journal of Psychological Assessment, 13 (2), pp. 140-152.

Camara, W. J. & Lane S. (2006). A historical perspective and current views on the Standards for Educational and Psychological Testing. Educational Measurement: Issues and Practice, 25 (3). 35-45.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. Educational Measurement: Issues and Practices, 20 (4), 19-27.

CRESST (Winter, 2002). Standards for educational accountability systems. Policy Brief 5. Los Angeles, CA: Author.

Joint Committee on Testing Practices (2004). Code of Fair Testing Practices in Education. Retrieved November 1, 2007, from <http://www.apa.org/science/FinalCode.pdf>

Kortz, D. (2006). Steps toward more effective implementation of the Standards for Educational and Psychological Testing. Educational Measurement: Issues and Practice, 25 (3). 46-50.

Madaus, G. F., Lynch, C. A., & Lynch, P. S. (2001). A brief history of attempts to monitor testing. National Board of Educational Testing and Public Policy Statements, 2 (2). Boston, MA: NBETPP.

Messick, S. (1989). Validity. In R. L. Linn (ed.). Educational measurement (3<sup>rd</sup> Ed.) (pp. 13-103). Washington, D.C.: American Council on Education-Macmillan.

National Association of Collegiate Admissions Counseling (2001). Statement of Principles of Good Practice. Alexandria, VA: Author.

National Association of School Psychologists (1997). Procedural guidelines for the adjudications of ethical complaints. Bethesda, MD: Author.

National Association of School Psychologists (4<sup>th</sup> Edition) (2000). Professional Conduct Manual. Bethesda, MD: Author.

National Council of Measurement in Education (1995). Code of professional responsibilities in education. Washington, DC: Author.

Novick, M. R. (1981). Federal guidelines and professional standards. American Psychologist, 36, 1035-1046.

Phillips, S. E. & Camara, W. J. (2006). Legal and ethical issues. In Brennan, R. (ed.) Educational measurement (4<sup>th</sup> Ed.) (pp 734-755.).

Rhoades, K. & Madaus, G. (May, 2003). Errors in standardized testing: A systematic problem. Retrieved November 1, 2007, from <http://www.bc.edu/research/nbetpp/statements/M1N4.pdf>.

Schmeiser, C. B. (1992). Ethical codes in the professions. Educational Measurement: Issues and Practice, 11 (3), 5-11.

Wise, L. (2006). Encouraging and supporting compliance with standards for educational testing. Educational Measurement: Issues and Practice, 25 (3). 51-56.