INTRODUCTION

Educational and psychological testing and assessment are among the most important contributions of cognitive and behavioral sciences to our society, providing fundamental and significant sources of information about individuals and groups. Not all tests are well developed, nor are all testing practices wise or beneficial, but there is extensive evidence documenting the usefulness of well-constructed, well-interpreted tests. Well-constructed tests that are valid for their intended purposes have the potential to provide substantial benefits for test takers and test users. Their proper use can result in better decisions about individuals and programs than would result without their use and can also provide a route to broader and more equitable access to education and employment. The improper use of tests, on the other hand, can cause considerable harm to test takers and other parties affected by test-based decisions. The intent of the Standards for Educational and Psychological Testing is to promote sound testing practices and to provide a basis for evaluating the quality of those practices. The Standards is intended for professionals who specify, develop, or select tests and for those who interpret, or evaluate the technical quality of, test results.

The Purpose of the Standards

The purpose of the Standards is to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses. Although such evaluations should depend heavily on professional judgment, the Standards provides a frame of reference to ensure that relevant issues are addressed. All professional test developers, sponsors, publishers, and users should make reasonable efforts to satisfy and follow the Standards and should encourage others to do so. All applicable standards should be met by all tests and in all test uses unless a sound professional reason is available to show why a standard is not relevant or technically feasible in a particular case.

The Standards makes no attempt to provide psychometric answers to questions of public policy regarding the use of tests. In general, the Standards advocates that, within feasible limits, the relevant technical information be made available so that those involved in policy decisions may be fully informed.

Legal Disclaimer

The Standards is not a statement of legal requirements, and compliance with the Standards is not a substitute for legal advice. Numerous federal, state, and local statutes, regulations, rules, and judicial decisions relate to some aspects of the use, production, maintenance, and development of tests and test results and impose standards that may be different for different types of testing. A review of these legal issues is beyond the scope of the Standards, the distinct purpose of which is to set forth the criteria for sound testing practices from the perspective of cognitive and behavioral science professionals. Where it appears that one or more standards address an issue on which established legal requirements may be particularly relevant, the standard, comment, or introductory material may make note of that fact. Lack of specific reference to legal requirements, however, does not imply the absence of a relevant legal requirement. When applying standards across international borders, legal differences may raise additional issues or require different treatment of issues.

In some areas, such as the collection, analysis, and use of test data and results for different subgroups, the law may both require participants in the testing process to take certain actions and prohibit those participants from taking other actions. Furthermore, because the science of testing is an evolving discipline, recent revisions to the Standards may not be reflected in existing legal authorities, including judicial decisions and agency
guidelines. In all situations, participants in the testing process should obtain the advice of counsel concerning applicable legal requirements.

In addition, although the Standards is not enforceable by the sponsoring organizations, it has been repeatedly recognized by regulatory authorities and courts as setting forth the generally accepted professional standards that developers and users of tests and other selection procedures follow. Compliance or noncompliance with the Standards may be used as relevant evidence of legal liability in judicial and regulatory proceedings. The Standards therefore merits careful consideration by all participants in the testing process.

Nothing in the Standards is meant to constitute legal advice. Moreover, the publishers disclaim any and all responsibility for liability created by participation in the testing process.

Tests and Test Uses to Which These Standards Apply

A test is a device or procedure in which a sample of an examinee’s behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process. Whereas the label test is sometimes reserved for instruments on which responses are evaluated for their correctness or quality, and the terms scale and inventory are used for measures of attitudes, interest, and dispositions, the Standards uses the single term test to refer to all such evaluative devices.

A distinction is sometimes made between tests and assessments. Assessment is a broader term than test, commonly referring to a process that integrates test information with information from other sources (e.g., information from other tests, inventories, and interviews; or the individual’s social, educational, employment, health, or psychological history). The applicability of the Standards to an evaluation device or method is determined by substance and not altered by the label applied to it (e.g., test, assessment, scale, inventory). The Standards should not be used as a checklist, as is emphasized in the section “Cautions to Be Considered in Using the Standards” at the end of this chapter.

Tests differ on a number of dimensions: the mode in which test materials are presented (e.g., paper-and-pencil, oral, or computerized administration); the degree to which stimulus materials are standardized; the type of response format (selection of a response from a set of alternatives, as opposed to the production of a free-form response); and the degree to which test materials are designed to reflect or simulate a particular context. In all cases, however, tests standardize the process by which test takers’ responses to test materials are evaluated and scored. As noted in prior versions of the Standards, the same general types of information are needed to judge the soundness of results obtained from using all varieties of tests.

The precise demarcation between measurement devices used in the fields of educational and psychological testing that do and do not fall within the purview of the Standards is difficult to identify. Although the Standards applies most directly to standardized measures generally recognized as “tests,” such as measures of ability, aptitude, achievement, attitudes, interests, personality, cognitive functioning, and mental health, the Standards may also be usefully applied in varying degrees to a broad range of less formal assessment techniques. Rigorous application of the Standards to unstandardized employment assessments (such as some job interviews) or to the broad range of unstructured behavior samples used in some forms of clinical and school-based psychological assessment (e.g., an intake interview), or to instructor-made tests that are used to evaluate student performance in education and training, is generally not possible. It is useful to distinguish between devices that lay claim to the concepts and techniques of the field of educational and psychological testing and devices that represent unstandardized or less standardized aids to day-to-day evaluative decisions. Although the principles and concepts underlying the Standards can be fruitfully applied to day-to-day decisions—such as when a business owner interviews a job applicant, a manager evaluates the performance of subordinates, a teacher develops a classroom assessment to monitor student progress toward an educational goal, or a coach evaluates a prospective athlete—it would be overreaching to
expect that the standards of the educational and psychological testing field be followed by those making such decisions. In contrast, a structured interviewing system developed by a psychologist and accompanied by claims that the system has been found to be predictive of job performance in a variety of other settings falls within the purview of the Standards. Adhering to the Standards becomes more critical as the stakes for the test taker and the need to protect the public increase.

**Participants in the Testing Process**

Educational and psychological testing and assessment involve and significantly affect individuals, institutions, and society as a whole. The individuals affected include students, parents, families, teachers, educational administrators, job applicants, employees, clients, patients, supervisors, executives, and evaluators, among others. The institutions affected include schools, colleges, businesses, industry, psychological clinics, and government agencies. Individuals and institutions benefit when testing helps them achieve their goals. Society, in turn, benefits when testing contributes to the achievement of individual and institutional goals.

There are many participants in the testing process, including, among others, (a) those who prepare and develop the test; (b) those who publish and market the test; (c) those who administer and score the test; (d) those who interpret test results for clients; (e) those who use the test results for some decision-making purpose (including policy makers and those who use data to inform social policy); (f) those who take the test by choice, direction, or necessity; (g) those who sponsor tests, such as boards that represent institutions or governmental agencies that contract with a test developer for a specific instrument or service; and (h) those who select or review tests, evaluating their comparative merits or suitability for the uses proposed. In general, those who are participants in the testing process should have appropriate knowledge of tests and assessments to allow them to make good decisions about which tests to use and how to interpret test results.

The interests of the various parties involved in the testing process may or may not be congruent. For example, when a test is given for counseling purposes or for job placement, the interests of the individual and the institution often coincide. In contrast, when a test is used to select from among many individuals for a highly competitive job or for entry into an educational or training program, the preferences of an applicant may be inconsistent with those of an employer or admissions officer. Similarly, when testing is mandated by a court, the interests of the test taker may be different from those of the party requesting the court order.

Individuals or institutions may serve several roles in the testing process. For example, in clinics the test taker is typically the intended beneficiary of the test results. In some situations the test administrator is an agent of the test developer, and sometimes the test administrator is also the test user. When an organization prepares its own employment tests, it is both the developer and the user. Sometimes a test is developed by a test author but published, marketed, and distributed by an independent publisher, although the publisher may play an active role in the test development process. Roles may also be further subdivided. For example, both an organization and a professional assessor may play a role in the provision of an assessment center. Given this intermingling of roles, it is often difficult to assign precise responsibility for addressing various standards to specific participants in the testing process. Uses of tests and testing practices are improved to the extent that those involved have adequate levels of assessment literacy.

Tests are designed, developed, and used in a wide variety of ways. In some cases, they are developed and “published” for use outside the organization that produces them. In other cases, as with state educational assessments, they are designed by the state educational agency and developed by contractors for exclusive and often one-time use by the state and not really “published” at all. Throughout the Standards, we use the general term test developer, rather than the more specific term test publisher, to denote those involved in
INTRODUCTION

the design and development of tests across the full range of test development scenarios.

The Standards is based on the premise that effective testing and assessment require that all professionals in the testing process possess the knowledge, skills, and abilities necessary to fulfill their roles, as well as an awareness of personal and contextual factors that may influence the testing process. For example, test developers and those selecting tests and interpreting test results need adequate knowledge of psychometric principles such as validity and reliability. They also should obtain any appropriate supervised experience and legislatively mandated practice credentials that are required to perform competently those aspects of the testing process in which they engage. All professionals in the testing process should follow the ethical guidelines of their profession.

Scope of the Revision

This volume serves as a revision of the 1999 Standards for Educational and Psychological Testing. The revision process started with the appointment of a Management Committee, composed of representatives of the three sponsoring organizations responsible for overseeing the general direction of the effort: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). To guide the revision, the Management Committee solicited and synthesized comments on the 1999 Standards from members of the sponsoring organizations and convened the Joint Committee for the Revision of the 1999 Standards in 2009 to do the actual revision. The Joint Committee also was composed of members of the three sponsoring organizations and was charged by the Management Committee with addressing five major areas: considering the accountability issues for use of tests in educational policy; broadening the concept of accessibility of tests for all examinees; representing more comprehensively the role of tests in the workplace; broadening the role of technology in testing; and providing for a better organizational structure for communicating the standards.

To be responsive to this charge, several actions were taken:

- The chapters “Educational Testing and Assessment” and “Testing in Program Evaluation and Public Policy,” in the 1999 version, were rewritten to attend to the issues associated with the uses of tests for educational accountability purposes.
- A new chapter, “Fairness in Testing,” was written to emphasize accessibility and fairness as fundamental issues in testing. Specific concerns for fairness are threaded throughout all of the chapters of the Standards.
- The chapter “Testing in Employment and Credentialing” (now “Workplace Testing and Credentialing”) was reorganized to more clearly identify when a standard is relevant to employment and/or credentialing.
- The impact of technology was considered throughout the volume. One of the major technology issues identified was the tension between the use of proprietary algorithms and the need for test users to be able to evaluate complex applications in areas such as automated scoring of essays, administering and scoring of innovative item types, and computer-based testing. These issues are considered in the chapter “Test Design and Development.”
- A content editor was engaged to help with the technical accuracy and clarity of each chapter and with consistency of language across chapters. As noted below, chapters in Part I (“Foundations”) and Part II (“Operations”) now have an “overarching standard” as well as themes under which the individual standards are organized. In addition, the glossary from the 1999 Standards for Educational and Psychological Testing was updated. As stated above, a major change in the organization of this volume involves the conceptualization of fairness. The 1999 edition had a part devoted to this topic, with separate chapters titled “Fairness in Testing and Test Use,” “Testing Individuals of Diverse Linguistic Backgrounds,” and “Testing Indi-
individually With Disabilities.” In the present edition, the topics addressed in those chapters are combined into a single, comprehensive chapter, and the chapter is located in Part I. This change was made to emphasize that fairness demands that all test takers be treated equitably. Fairness and accessibility, the unobstructed opportunity for all examinees to demonstrate their standing on the construct(s) being measured, are relevant for valid score interpretations for all individuals and subgroups in the intended population of test takers. Because issues related to fairness in testing are not restricted to individuals with diverse linguistic backgrounds or those with disabilities, the chapter was more broadly cast to support appropriate testing experiences for all individuals. Although the examples in the chapter often refer to individuals with diverse linguistic and cultural backgrounds and individuals with disabilities, they also include examples relevant to gender and to older adults, people of various ethnicities and racial backgrounds, and young children, to illustrate potential barriers to fair and equitable assessment for all examinees.

**Organization of the Volume**

Part I of the Standards, “Foundations,” contains standards for validity (chap. 1); reliability/precision and errors of measurement (chap. 2); and fairness in testing (chap. 3). Part II, “Operations,” addresses test design and development (chap. 4); scores, scales, norms, score linking, and cut scores (chap. 5); test administration, scoring, reporting, and interpretation (chap. 6); supporting documentation for tests (chap. 7); the rights and responsibilities of test takers (chap. 8); and the rights and responsibilities of test users (chap. 9). Part III, “Testing Applications,” treats specific applications in psychological testing and assessment (chap. 10); workplace testing and credentialing (chap. 11); educational testing and assessment (chap. 12); and uses of tests for program evaluation, policy studies, and accountability (chap. 13). Also included is a glossary, which provides definitions for terms as they are used specifically in this volume.

Each chapter begins with introductory text that provides background for the standards that follow. Although the introductory text is at times prescriptive, it should not be interpreted as imposing additional standards.

**Categories of Standards**

The text of each standard and any accompanying commentary include the conditions under which a standard is relevant. Depending on the context and purpose of test development or use, some standards will be more salient than others. Moreover, some standards are broad in scope, setting forth concerns or requirements relevant to nearly all tests or testing contexts, and other standards are narrower in scope. However, all standards are important in the contexts to which they apply. Any classification that gives the appearance of elevating the general importance of some standards over others could invite neglect of certain standards that need to be addressed in particular situations. Rather than differentiate standards using priority labels, such as “primary,” “secondary,” or “conditional” (as were used in the 1985 Standards), this edition emphasizes that unless a standard is deemed clearly irrelevant, inappropriate, or technically infeasible for a particular use, all standards should be met, making all of them essentially “primary” for that context.

Unless otherwise specified in a standard or commentary, and with the caveats outlined below, standards should be met before operational test use. Each standard should be carefully considered to determine its applicability to the testing context under consideration. In a given case there may be a sound professional reason that adherence to the standard is inappropriate. There may also be occasions when technical feasibility influences whether a standard can be met prior to operational test use. For example, some standards may call for analyses of data that are not available at the point of initial operational test use. In other cases, traditional quantitative analyses may not be feasible due to small sample sizes. However, there may be other methodologies that could be used to gather information to support the standard, such as small sample methodologies, qualitative
studies, focus groups, and even logical analysis. In such instances, test developers and users should make a good faith effort to provide the kinds of data called for in the standard to support the valid interpretations of the test results for their intended purposes. If test developers, users, and, when applicable, sponsors have deemed a standard to be inapplicable or technically infeasible, they should be able, if called upon, to explain the basis for their decision. However, there is no expectation that documentation of all such decisions be routinely available.

**Presentation of Individual Standards**

Individual standards are presented after an introductory text that presents some key concepts for interpreting and applying the standards. In many cases, the standards themselves are coupled with one or more comments. These comments are intended to amplify, clarify, or provide examples to aid in the interpretation of the meaning of the standards. The standards often direct a developer or user to implement certain actions. Depending on the type of test, it is sometimes not clear in the statement of a standard to whom the standard is directed. For example, Standard 1.2 in the chapter "Validity" states:

A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.

The party responsible for implementing this standard is the party or person who is articulating the recommended interpretation of the test scores. This may be a test user, a test developer, or someone who is planning to use the test scores for a particular purpose, such as making classification or licensure decisions. It often is not possible in the statement of a standard to specify who is responsible for such actions; it is intended that the party or person performing the action specified in the standard be the party responsible for adhering to the standard.

Some of the individual standards and introductory text refer to groups and subgroups. The term group is generally used to identify the full examinee population, referred to as the intended examinee group, the intended test-taker group, the intended examinee population, or the population. A subgroup includes members of the larger group who are identifiable in some way that is relevant to the standard being applied. When data or analyses are indicated for various subgroups, they are generally referred to as subgroups within the intended examinee group, groups from the intended examinee population, or relevant subgroups.

In applying the Standards, it is important to bear in mind that the intended referent subgroups for the individual standards are context specific. For example, referent ethnic subgroups to be considered during the design phase of a test would depend on the expected ethnic composition of the intended test group. In addition, many more subgroups could be relevant to a standard dealing with the design of fair test questions than to a standard dealing with adaptations of a test's format. Users of the Standards will need to exercise professional judgment when deciding which particular subgroups are relevant for the application of a specific standard.

In deciding which subgroups are relevant for a particular standard, the following factors, among others, may be considered: credible evidence that suggests a group may face particular construct-irrelevant barriers to test performance, statutes or regulations that designate a group as relevant to score interpretations, and large numbers of individuals in the group within the general population. Depending on the context, relevant subgroups might include, for example, males and females, individuals of differing socioeconomic status, individuals differing by race and/or ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds (particularly when testing extends across international borders), individuals with disabilities, young children, or older adults.

Numerous examples are provided in the Standards to clarify points or to provide illustrations of how to apply a particular standard. Many of
the examples are drawn from research with students with disabilities or persons from diverse language or cultural groups; fewer, from research with other identifiable groups, such as young children or adults. There was also a purposeful effort to provide examples for educational, psychological, and industrial settings.

The standards in each chapter in Parts I and II (“Foundations” and “Operations”) are introduced by an overarching standard, designed to convey the central intent of the chapter. These overarching standards are always numbered with .0 following the chapter number. For example, the overarching standard in chapter 1 is numbered 1.0. The overarching standards summarize guiding principles that are applicable to all tests and test uses. Further, the themes and standards in each chapter are ordered to be consistent with the sequence of the material in the introductory text for the chapter. Because some users of the Standards may turn only to chapters directly relevant to a given application, certain standards are repeated in different chapters, particularly in Part III, “Testing Applications.” When such repetition occurs, the essence of the standard is the same. Only the wording, area of application, or level of elaboration in the comment is changed.

Cautions to Be Considered in Using the Standards

In addition to the legal disclaimer set forth above, several cautions are important if we are to avoid misinterpretations, misapplications, and misuses of the Standards:

- Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and the acceptability of a test or test application cannot be determined by using a checklist. Specific circumstances affect the importance of individual standards, and individual standards should not be considered in isolation. Therefore, evaluating acceptability depends on (a) professional judgment that is based on a knowledge of behavioral science, psychometrics, and the relevant standards in the professional field to which the test applies; (b) the degree to which the intent of the standard has been satisfied by the test developer and user; (c) the alternative measurement devices that are readily available; (d) research and experiential evidence regarding the feasibility of meeting the standard; and (e) applicable laws and regulations.

- When tests are at issue in legal proceedings and other situations requiring expert witness testimony, it is essential that professional judgment be based on the accepted corpus of knowledge in determining the relevance of particular standards in a given situation. The intent of the Standards is to offer guidance for such judgments.

- Claims by test developers or test users that a test, manual, or procedure satisfies or follows the standards in this volume should be made with care. It is appropriate for developers or users to state that efforts were made to adhere to the Standards, and to provide documents describing and supporting those efforts. Blanket claims without supporting evidence should not be made.

- The standards are concerned with a field that is rapidly evolving. Consequently, there is a continuing need to monitor changes in the field and to revise this document as knowledge develops. The use of older versions of the Standards may be a disservice to test users and test takers.

- Requiring the use of specific technical methods is not the intent of the Standards. For example, where specific statistical reporting requirements are mentioned, the phrase “or generally accepted equivalent” should always be understood.