EXHIBIT 70

STANDARDS
for Educational and Psychological Testing

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
AMERICAN PSYCHOLOGICAL ASSOCIATION
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Published by the
American Educational Research Association
1430 K St., NW, Suite 1200
Washington, DC 20005

Printed in the United States of America

Prepared by the
Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education

Library of Congress Cataloging-in-Publication Data

American Educational Research Association.
Standards for educational and psychological testing / American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
pages cm
"Prepared by the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association and National Council on Measurement in Education"—T.p. verso.
Include index.
ISBN 978-0-935302-35-6 (alk. paper)
1. Educational tests and measurements—Standards—United States. 2. Psychological tests—Standards—United States. I. American Psychological Association. II. National Council on Measurement in Education. III. Joint Committee on Standards for Educational and Psychological Testing (U.S.) IV. Title.
LB3051.A693 2014
371.26'0973—dc23

2014009333

CONTE

CONTENTS

PREFA

This edition of
chological Testi
Educational R
American Psyc
the National Co
(NCME). Earl
organizations
use of tests. The
for Psychologic
prepared by an
APA in 1954.
mendations for
committee rep
Council on M
(NCMUE) an
cation Associa

The third,
prepared by a j
APA, and NC
1966. It was t
Educational an
as the *Standa*
the *Standards*
representing A
in 1974, 1985

The curren
was formed b
three sponsorin
of one represen
committee's re
whether the 1
then creating th
for a joint com
co-chairs and
development
related to the
Standards.

Standards Ma

Wayne J. Cama
David Frisbie (2
Suzanne Lane, a
Barbara S. Plak

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

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-

viduals 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.

STANDARDS FOR FAIRNESS

The standards in this chapter begin with an overarching standard (numbered 3.0), which is designed to convey the central intent or primary focus of the chapter. The overarching standard may also be viewed as the guiding principle of the chapter, and is applicable to all tests and test users. All subsequent standards have been separated into four thematic clusters labeled as follows:

1. Test Design, Development, Administration, and Scoring Procedures That Minimize Barriers to Valid Score Interpretations for the Widest Possible Range of Individuals and Relevant Subgroups
2. Validity of Test Score Interpretations for Intended Uses for the Intended Examinee Population
3. Accommodations to Remove Construct-Irrelevant Barriers and Support Valid Interpretations of Scores for Their Intended Uses
4. Safeguards Against Inappropriate Score Interpretations for Intended Uses

Standard 3.0

All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.

Comment: The central idea of fairness in testing is to identify and remove construct-irrelevant barriers to maximal performance for any examinee. Removing these barriers allows for the comparable and valid interpretation of test scores for all examinees. Fairness is thus central to the validity and comparability of the interpretation of test scores for intended uses.

Cluster 1. Test Design, Development, Administration, and Scoring Procedures That Minimize Barriers to Valid Score Interpretations for the Widest Possible Range of Individuals and Relevant Subgroups

Standard 3.1

Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.

Comment: Test developers must clearly delineate both the constructs that are to be measured by the test and the characteristics of the individuals and subgroups in the intended population of test takers. Test tasks and items should be designed to maximize access and be free of construct-irrelevant barriers as far as possible for all individuals and relevant subgroups in the intended test-taker population. One way to accomplish these goals is to create the test using principles of universal design, which take account of the characteristics of all individuals for whom the test is intended and include such elements as precisely defining constructs and avoiding, where possible, characteristics and formats of items and tests (for example, test speededness) that may compromise valid score interpretations for individuals or relevant subgroups. Another principle of universal design is to provide simple, clear, and intuitive testing procedures and instructions. Ultimately, the goal is to design a testing process that will, to the extent practicable, remove potential barriers to the measurement of the intended construct for all individuals, including those individuals requiring accommodations. Test developers need to be knowledgeable about group differences that may interfere

with the precision of scores and the validity of test score inferences, and they need to be able to take steps to reduce bias.

Standard 3.2

Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

Comment: Unnecessary linguistic, communicative, cognitive, cultural, physical, and/or other characteristics in test item stimulus and/or response requirements can impede some individuals in demonstrating their standing on intended constructs. Test developers should use language in tests that is consistent with the purposes of the tests and that is familiar to as wide a range of test takers as possible. Avoiding the use of language that has different meanings or different connotations for relevant subgroups of test takers will help ensure that test takers who have the skills being assessed are able to understand what is being asked of them and respond appropriately. The level of language proficiency, physical response, or other demands required by the test should be kept to the minimum required to meet work and credentialing requirements and/or to represent the target construct(s). In work situations, the modality in which language proficiency is assessed should be comparable to that required on the job, for example, oral and/or written, comprehension and/or production. Similarly, the physical and verbal demands of response requirements should be consistent with the intended construct.

Standard 3.3

Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test.

Comment: Test developers should include individuals from relevant subgroups of the intended testing population in pilot or field test samples used to evaluate item and test appropriateness for construct interpretations. The analyses that are carried out using pilot and field testing data should seek to detect aspects of test design, content, and format that might distort test score interpretations for the intended uses of the test scores for particular groups and individuals. Such analyses could employ a range of methodologies, including those appropriate for small sample sizes, such as expert judgment, focus groups, and cognitive labs. Both qualitative and quantitative sources of evidence are important in evaluating whether items are psychometrically sound and appropriate for all relevant subgroups.

If sample sizes permit, it is often valuable to carry out separate analyses for relevant subgroups of the population. When it is not possible to include sufficient numbers in pilot and/or field test samples in order to do separate analyses, operational test results may be accumulated and used to conduct such analyses when sample sizes become large enough to support the analyses.

If pilot or field test results indicate that items or tests function differentially for individuals from, for example, relevant age, cultural, disability, gender, linguistic and/or racial/ethnic groups in the population of test takers, test developers should investigate aspects of test design, content, and format (including response formats) that might contribute to the differential performance of members of these groups and, if warranted, eliminate these aspects from future test development practices.

Expert and sensitivity reviews can serve to guard against construct-irrelevant language and images, including those that may offend some individuals or subgroups, and against construct-irrelevant context that may be more familiar to some than others. Test publishers often conduct sensitivity reviews of all test material to detect and remove sensitive material from tests (e.g., text, graphics, and other visual representations within the test that could be seen as offensive to some groups and possibly affect the scores of individuals from these groups). Such reviews should be conducted before a test becomes operational.

Standard 3.4

Test takers shoul
during the test ad

Comment: Thos
adhere to standard
and security pro
reflect the const
not be unduly inf
testing process. T
istration should m
predispositions th
istration or interp

Computerized
gy-based testing a
ization in adminis
must have access
the technology its
aminees working
be unfairly disadva
on newer equipme
differ in speed of
one screen to the
visuals, or in other
that construct-irre
test performance.

Issues related
administration can
of treatment of in
fairness of test scor
unauthorized distr
aminees but not o
ministrations whe
ensured, could pro
takers over others.
should be interpret

Standard 3.5

Test developers sh
provisions that hav
tration and scoring
struct-irrelevant bar
in the test-taker po

Comment: Test de
construct-irrelevant

Standard 3.4

Test takers should receive comparable treatment during the test administration and scoring process.

Comment: Those responsible for testing should adhere to standardized test administration, scoring, and security protocols so that test scores will reflect the construct(s) being assessed and will not be unduly influenced by idiosyncrasies in the testing process. Those responsible for test administration should mitigate the possibility of personal predispositions that might affect the test administration or interpretation of scores.

Computerized and other forms of technology-based testing add extra concerns for standardization in administration and scoring. Examinees must have access to technology so that aspects of the technology itself do not influence scores. Examinees working on older, slower equipment may be unfairly disadvantaged relative to those working on newer equipment. If computers or other devices differ in speed of processing or movement from one screen to the next, in the fidelity of the visuals, or in other important ways, it is possible that construct-irrelevant factors may influence test performance.

Issues related to test security and fidelity of administration can also threaten the comparability of treatment of individuals and the validity and fairness of test score interpretations. For example, unauthorized distribution of items to some examinees but not others, or unproctored test administrations where standardization cannot be ensured, could provide an advantage to some test takers over others. In these situations, test results should be interpreted with caution.

Standard 3.5

Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population.

Comment: Test developers should specify how construct-irrelevant barriers were minimized in the test development process for individuals from all relevant subgroups in the intended test population. Test developers and/or users should also document any studies carried out to examine the reliability/precision of scores and validity of scorer interpretations for relevant subgroups of the intended population of test takers for the intended uses of the test scores. Special test administration, scoring, and reporting procedures should be documented and made available to test users.

Cluster 2. Validity of Test Score Interpretations for Intended Uses for the Intended Examinee Population

Standard 3.6

Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws.

Comment: Subgroup mean differences do not in and of themselves indicate lack of fairness, but such differences should trigger follow-up studies, where feasible, to identify the potential causes of such differences. Depending on whether subgroup differences are discovered during the development or use phase, either the test developer or the test user is responsible for initiating follow-up inquiries and, as appropriate, relevant studies. The inquiry should investigate construct underrepresentation and sources of construct-irrelevant variance as potential causes of subgroup differences, investigated as feasible, through quantitative and/or qualitative studies. The kinds of validity evidence considered may include analysis of test content, internal structure of test responses, the relationship of test scores to other variables, or the response processes employed by the individual examinees. When

sample sizes are sufficient, studies of score precision and accuracy for relevant subgroups also should be conducted. When sample sizes are small, data may sometimes be accumulated over operational administrations of the test so that suitable quantitative analyses by subgroup can be performed after the test has been in use for a period of time. Qualitative studies also are relevant to the supporting validity arguments (e.g., expert reviews, focus groups, cognitive labs). Test developers should closely consider findings from quantitative and/or qualitative analyses in documenting the interpretations for the intended score uses, as well as in subsequent test revisions.

Analyses, where possible, may need to take into account the level of heterogeneity within relevant subgroups, for example, individuals with different disabilities, or linguistic minority examinees at different levels of English proficiency. Differences within these subgroups may influence the appropriateness of test content, the internal structure of the test responses, the relation of test scores to other variables, or the response processes employed by individual examinees.

Standard 3.7

When criterion-related validity evidence is used as a basis for test score–based predictions of future performance and sample sizes are sufficient, test developers and/or users are responsible for evaluating the possibility of differential prediction for relevant subgroups for which there is prior evidence or theory suggesting differential prediction.

Comment: When sample sizes are sufficient, differential prediction is often examined using regression analysis. One approach to regression analysis examines slope and intercept differences between targeted groups (e.g., Black and White samples), while another examines systematic deviations from a common regression line for the groups of interest. Both approaches can account for the possibility of predictive bias and/or differences in heterogeneity between groups and provide valuable information for the examination of differential predictions. In contrast, correlation coefficients provide inadequate evidence for or against a differential prediction hypothesis if groups or treatments are found to have unequal means and variances on the test and the criterion. It is particularly important in the context of testing for high-stakes purposes that test developers and/or users examine differential prediction and avoid the use of correlation coefficients in situations where groups or treatments result in unequal means or variances on the test and criterion.

Standard 3.8

When tests require the scoring of constructed responses, test developers and/or users should collect and report evidence of the validity of score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores.

Comment: Subgroup differences in examinee responses and/or the expectations and perceptions of scorers can introduce construct-irrelevant variance in scores from constructed response tests. These, in turn, could seriously affect the reliability/precision, validity, and comparability of score interpretations for intended uses for some individuals. Different methods of scoring could differentially influence the construct representation of scores for individuals from some subgroups.

For human scoring, scoring procedures should be designed with the intent that the scores reflect the examinee's standing relative to the tested construct(s) and are not influenced by the perceptions and personal predispositions of the scorers. It is essential that adequate training and calibration of scorers be carried out and monitored throughout the scoring process to support the consistency of scorers' ratings for individuals from relevant subgroups. Where sample sizes permit, the precision and accuracy of scores for relevant subgroups also should be calculated.

Automated scoring algorithms may be used to score complex constructed responses, such as essays, either as the sole determiner of the score or in conjunction with a score provided by a human

scorer. Scoring algorithms need to be reviewed for potential sources of bias. The precision of scores and validity of score interpretations resulting from automated scoring should be evaluated for all relevant subgroups of the intended population.

Cluster 3. Accommodations to Remove Construct-Irrelevant Barriers and Support Valid Interpretations of Scores for Their Intended Uses

Standard 3.9

Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

Comment: Test accommodations are designed to remove construct-irrelevant barriers related to individual characteristics that otherwise would interfere with the measurement of the target construct and therefore would unfairly disadvantage individuals with these characteristics. These accommodations include changes in administration setting, presentation, interface/engagement, and response requirements, and may include the addition of individuals to the administration process (e.g., readers, scribes).

An appropriate accommodation is one that responds to specific individual characteristics but does so in a way that does not change the construct the test is measuring or the meaning of scores. Test developers and/or test users should document the basis for the conclusion that the accommodation does not change the construct that the test is measuring. Accommodations must address individual test takers' specific needs (e.g., cognitive, linguistic, sensory, physical) and may be required by law. For example, individuals who are not fully proficient in English may need linguistic accommodations that address their language status, while visually impaired individuals may need text magnification. In many cases when a test is used to evaluate the academic progress of an individual, the accommodation that will best eliminate construct irrelevance will match the accommodation used for instruction.

Test modifications that change the construct that the test is measuring may be needed for some examinees to demonstrate their standing on some aspect of the intended construct. If an assessment is modified to improve access to the intended construct for designated individuals, the modified assessment should be treated like a newly developed assessment that needs to adhere to the test standards for validity, reliability/precision, fairness, and so forth.

Standard 3.10

When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.

Comment: Test accommodations should be used only when the test taker has a documented need for the accommodation, for example, an Individualized Education Plan (IEP) or documentation by a physician, psychologist, or other qualified professional. The documentation should be prepared in advance of the test-taking experience and reviewed by one or more experts qualified to make a decision about the relevance of the documentation to the requested accommodation.

Test developers and/or users should provide individuals requiring accommodations in a testing situation with information about the availability of accommodations and the procedures for requesting them prior to the test administration. In settings where accommodations are routinely provided for individuals with documented needs (e.g., educational settings), the documentation should describe permissible accommodations and include standardized protocols and/or procedures for identifying examinees eligible for accommodations, identifying and assigning appropriate accommodations for these individuals, and administering accommodations, scoring, and reporting in accordance with standardized rules.

Test administrators and users should also provide those who have a role in determining and administering accommodations with sufficient information and expertise to appropriately use accommodations that may be applied to the assessment. Instructions for administering any changes in the test or testing procedures should be clearly documented and, when necessary, test administrators should be trained to follow these procedures. The test administrator should administer the accommodations in a standardized manner as documented by the test developer. Administration procedures should include procedures for recording which accommodations were used for specific individuals and, where relevant, for recording any deviation from standardized procedures for administering the accommodations.

The test administrator or appropriate representative of the test user should document any use of accommodations. For large-scale education assessments, test users also should monitor the appropriate use of accommodations.

Standard 3.11

When a test is changed to remove barriers to the accessibility of the construct being measured, test developers and/or users are responsible for obtaining and documenting evidence of the validity of score interpretations for intended uses of the changed test, when sample sizes permit.

Comment: It is desirable, where feasible and appropriate, to pilot and/or field test any test alterations with individuals representing each relevant subgroup for whom the alteration is intended. Validity studies typically should investigate both the efficacy of the alteration for intended subgroup(s) and the comparability of score inferences from the altered and original tests.

In some circumstances, developers may not be able to obtain sufficient samples of individuals, for example, those with the same disability or similar levels of a disability, to conduct standard empirical analyses of reliability/precision and validity. In these situations, alternative ways should be sought to evaluate the validity of the changed test for relevant subgroups, for example through small-sample qualitative studies or professional judgments that examine the comparability of the original and altered tests and/or that investigate alternative explanations for performance on the changed tests.

Evidence should be provided for recommended alterations. If a test developer recommends different time limits, for example, for individuals with disabilities or those from diverse linguistic and cultural backgrounds, pilot or field testing should be used, whenever possible, to establish these particular time limits rather than simply allowing test takers a multiple of the standard time without examining the utility of the arbitrary implementation of multiples of the standard time. When possible, fatigue and other time-related issues should be investigated as potentially important factors when time limits are extended.

When tests are linguistically simplified to remove construct-irrelevant variance, test developers and/or users are responsible for documenting evidence of the comparability of scores from the linguistically simplified tests to the original test, when sample sizes permit.

Standard 3.12

When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretations for intended use.

Comment: The term *adaptation* is used here to describe changes made to tests translated from one language to another to reduce construct-irrelevant variance that may arise due to individual or subgroup characteristics. In this case the translation/adaptation process involves not only translating the language of the test so that it is suitable for the subgroup taking the test, but also addressing any construct-irrelevant linguistic and cultural subgroup characteristics that may interfere with

measurement of the intended construct(s). When multiple language versions of a test are intended to provide comparable scores, test developers should describe in detail the methods used for test translation and adaptation and should report evidence of test score validity pertinent to the linguistic and cultural groups for whom the test is intended and pertinent to the scores' intended uses. Evidence of validity may include empirical studies and/or professional judgment documenting that the different language versions measure comparable or similar constructs and that the score interpretations from the two versions have comparable validity for their intended uses. For example, if a test is translated and adapted into Spanish for use with Central American, Cuban, Mexican, Puerto Rican, South American, and Spanish populations, the validity of test score interpretations for specific uses should be evaluated with members of each of these groups separately, where feasible. Where sample sizes permit, evidence of score accuracy and precision should be provided for each group, and test properties for each subgroup should be included in test manuals.

Standard 3.13

A test should be administered in the language that is most relevant and appropriate to the test purpose.

Comment: Test users should take into account the linguistic and cultural characteristics and relative language proficiencies of examinees who are bilingual or use multiple languages. Identifying the most appropriate language(s) for testing also requires close consideration of the context and purpose for testing. Except in cases where the purpose of testing is to determine test takers' level of proficiency in a particular language, the test takers should be tested in the language in which they are most proficient. In some cases, test takers' most proficient language in general may not be the language in which they were instructed or trained in relation to tested constructs, and in these cases it may be more appropriate to administer the test in the language of instruction.

Professional judgment needs to be used to determine the most appropriate procedures for establishing relative language proficiencies. Such procedures may range from self-identification by examinees to formal language proficiency testing. Sensitivity to linguistic and cultural characteristics may require the sole use of one language in testing or use of multiple languages to minimize the introduction of construct-irrelevant components into the measurement process.

Determination of a test taker's most proficient language for test administration does not automatically guarantee validity of score inferences for the intended use. For example, individuals may be more proficient in one language than another, but not necessarily developmentally proficient in either; disconnects between the language of construct acquisition and that of assessment also can compromise appropriate interpretation of the test taker's scores.

Standard 3.14

When testing requires the use of an interpreter, the interpreter should follow standardized procedures and, to the extent feasible, be sufficiently fluent in the language and content of the test and the examinee's native language and culture to translate the test and related testing materials and to explain the examinee's test responses, as necessary.

Comment: Although individuals with limited proficiency in the language of the test (including deaf and hard-of-hearing individuals whose native language may be sign language) should ideally be tested by professionally trained bilingual/bicultural examiners, the use of an interpreter may be necessary in some situations. If an interpreter is required, the test user is responsible for selecting an interpreter with reasonable qualifications, experience, and preparation to assist appropriately in the administration of the test. As with other aspects of standardized testing, procedures for administering a test when an interpreter is used should be standardized and documented. It is necessary for the interpreter to understand the

importance of following standardized procedures for this test, the importance of accurately conveying to the examiner an examinee's actual responses, and the role and responsibilities of the interpreter in testing. When the translation of technical terms is important to accurately assess the construct, the interpreter should be familiar with the meaning of these terms and corresponding vocabularies in the respective languages.

Unless a test has been standardized and normed with the use of interpreters, their use may need to be viewed as an alteration that could change the measurement of the intended construct, in particular because of the introduction of a third party during testing, as well as the modification of the standardized protocol. Differences in word meaning, familiarity, frequency, connotations, and associations make it difficult to directly compare scores from any non-standardized translations to English-language norms.

When a test is likely to require the use of interpreters, the test developer should provide clear guidance on how interpreters should be selected and their role in administration.

Cluster 4. Safeguards Against Inappropriate Score Interpretations for Intended Uses

Standard 3.15

Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

Comment: Test developers should include in test manuals and instructions for score interpretation explicit statements about the applicability of the test for relevant subgroups. Test developers should provide evidence of the applicability of the test for relevant subgroups and make explicit cautions against foreseeable (based on prior experience or other relevant sources such as research literature) misuses of test results.

Standard 3.16

When credible research indicates that test scores for some relevant subgroups are differentially affected by construct-irrelevant characteristics of the test or of the examinees, when legally permissible, test users should use the test only for those subgroups for which there is sufficient evidence of validity to support score interpretations for the intended uses.

Comment: A test may not measure the same construct(s) for individuals from different relevant subgroups because different characteristics of test content or format influence scores of test takers from one subgroup to another. Any such differences may inadvertently advantage or disadvantage individuals from these subgroups. The decision whether to use a test with any given relevant subgroup necessarily involves a careful analysis of the validity evidence for the subgroup, as is called for in Standard 1.4. The decision also requires consideration of applicable legal requirements and the exercise of thoughtful professional judgment regarding the significance of any construct-irrelevant components. In cases where there is credible evidence of differential validity, developers should provide clear guidance to the test user about when and whether valid interpretations of scores for their intended uses can or cannot be drawn for individuals from these subgroups.

There may be occasions when examinees request or demand to take a version of the test other than that deemed most appropriate by the developer or user. For example, an individual with a disability may decline an altered format and request the standard form. Acceding to such requests, after fully informing the examinee about the characteristics of the test, the accommodations that are available, and how the test scores will be used, is not a violation of this standard and in some instances may be required by law.

In some cases, such as when a test will distribute benefits or burdens (such as qualifying for an honors class or denial of a promotion in a job), the law may limit the extent to which a test user

may evaluate some groups under the test and other groups under a different test.

Standard 3.17

When aggregate scores are publicly reported for relevant subgroups—for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults—test users are responsible for providing evidence of comparability and for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups.

Comment: Reporting scores for relevant subgroups is justified only if the scores have comparable meaning across these groups and there is sufficient sample size per group to protect individual identity and warrant aggregation. This standard is intended to be applicable to settings where scores are implicitly or explicitly presented as comparable in meaning across subgroups. Care should be taken that the terms used to describe reported subgroups are clearly defined, consistent with common usage, and clearly understood by those interpreting test scores.

Terminology for describing specific subgroups for which valid test score inferences can and cannot be drawn should be as precise as possible, and categories should be consistent with the intended uses of the results. For example, the terms *Latino* or *Hispanic* can be ambiguous if not specifically defined, in that they may denote individuals of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish-culture origin, regardless of race/ethnicity, and may combine those who are recent immigrants with those who are U.S. native born, those who may not be proficient in English, and those of diverse socioeconomic background. Similarly, the term "individuals with disabilities" encompasses a wide range of specific conditions and background characteristics. Even references to specific categories of individuals with disabilities, such as hearing impaired, should be accompanied by an explanation of the meaning of the term and an indication of the variability of individuals within the group.

Standard 3.18

In testing individuals for diagnostic and/or special program placement purposes, test users should not use test scores as the sole indicators to characterize an individual's functioning, competence, attitudes, and/or predispositions. Instead, multiple sources of information should be used, alternative explanations for test performance should be considered, and the professional judgment of someone familiar with the test should be brought to bear on the decision.

Comment: Many test manuals point out variables that should be considered in interpreting test scores, such as clinically relevant history, medications, school record, vocational status, and test-taker motivation. Influences associated with variables such as age, culture, disability, gender, and linguistic or racial/ethnic characteristics may also be relevant.

Opportunity to learn is another variable that may need to be taken into account in educational and/or clinical settings. For instance, if recent immigrants being tested on a personality inventory or an ability measure have little prior exposure to school, they may not have had the opportunity to learn concepts that the test assumes are common knowledge or common experience, even if the test is administered in the native language. Not taking into account prior opportunity to learn can lead to misdiagnoses, inappropriate placements and/or services, and unintended negative consequences.

Inferences about test takers' general language proficiency should be based on tests that measure a range of language features, not a single linguistic skill. A more complete range of communicative abilities (e.g., word knowledge, syntax as well as cultural variation) will typically need to be assessed. Test users are responsible for interpreting individual

scores in light of alternative explanations and/or relevant individual variables noted in the test manual.

Standard 3.19

In settings where the same authority is responsible for both provision of curriculum and high-stakes decisions based on testing of examinees' curriculum mastery, examinees should not suffer permanent negative consequences if evidence indicates that they have not had the opportunity to learn the test content.

Comment: In educational settings, students' opportunity to learn the content and skills assessed by an achievement test can seriously affect their test performance and the validity of test score interpretations for intended use for high-stakes individual decisions. If there is not a good match between the content of curriculum and instruction and that of tested constructs for some students, those students cannot be expected to do well on the test and can be unfairly disadvantaged by high-stakes individual decisions, such as denying high school graduation, that are made based on test results. When an authority, such as a state or district, is responsible for prescribing and/or delivering curriculum and instruction, it should not penalize individuals for test performance on content that the authority has not provided.

Note that this standard is not applicable in situations where different authorities are responsible for curriculum, testing, and/or interpretation and use of results. For example, opportunity to learn may be beyond the knowledge or control of test users, and it may not influence the validity of test interpretations such as predictions of future performance.

Standard 3.20

When a construct can be measured in different ways that are equal in their degree of construct representation and validity (including freedom from construct-irrelevant variance), test users should consider, among other factors, evidence of subgroup differences in mean scores or in percentages of examinees whose scores exceed the cut scores, in deciding which test and/or cut scores to use.

Comment: Evidence of differential subgroup performance is one important factor influencing the choice between one test and another. However, other factors, such as cost, testing time, test security, and logistical issues (e.g., the need to screen very large numbers of examinees in a very short time), must also enter into professional judgments about test selection and use. If the scores from two tests lead to equally valid interpretations and impose similar costs or other burdens, legal considerations may require selecting the test that minimizes subgroup differences.