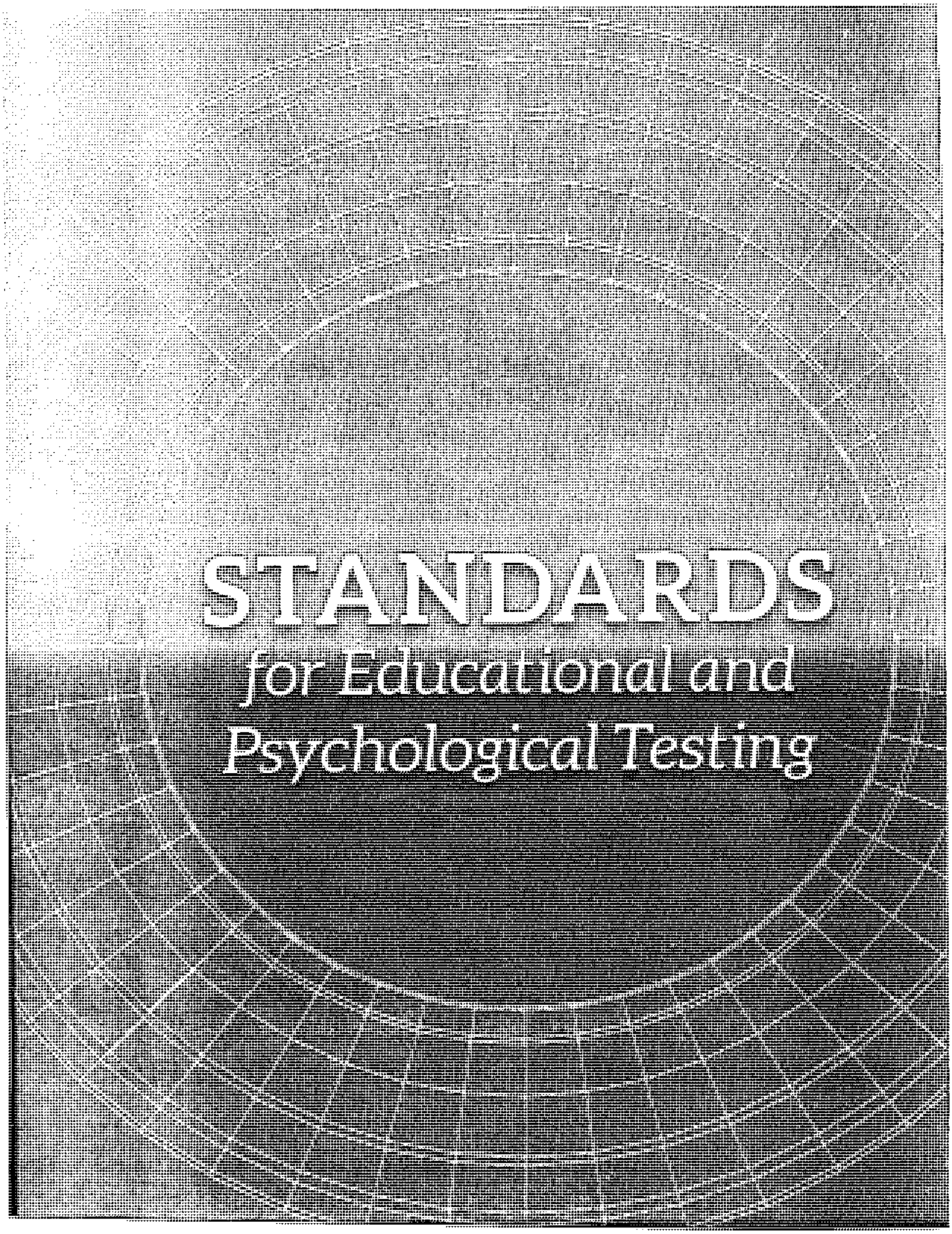

STANDARDS for Educational and Psychological Testing

CONTENTS

INTRODUCTION

PART I
FOUNDATIONS

PART II
OPERATIONS

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-

3. FAIRNESS IN TESTING

BACKGROUND

This chapter addresses the importance of fairness as a fundamental issue in protecting test takers and test users in all aspects of testing. The term *fairness* has no single technical meaning and is used in many different ways in public discourse. It is possible that individuals endorse fairness in testing as a desirable social goal, yet reach quite different conclusions about the fairness of a given testing program. A full consideration of the topic would explore the multiple functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. It would consider the technical properties of tests, the ways in which test results are reported and used, the factors that affect the validity of score interpretations, and the consequences of test use. A comprehensive analysis of fairness in testing also would examine the regulations, statutes, and case law that govern test use and the remedies for harmful testing practices. The *Standards* cannot hope to deal adequately with all of these broad issues, some of which have occasioned sharp disagreement among testing specialists and others interested in testing. Our focus must be limited here to delineating the aspects of tests, testing, and test use that relate to fairness as described in this chapter, which are the responsibility of those who develop, use, and interpret the results of tests, and upon which there is general professional and technical agreement.

Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use. In previous versions of the *Standards*, fairness and the assessment of individuals from specific subgroups of test takers, such as individuals with disabilities and individuals with diverse linguistic and cultural backgrounds, were presented in separate chapters. In the current version of the *Standards*, these issues are presented in a single chapter to emphasize that fairness to all individuals in the intended population of test takers is an overriding, foundational concern, and that common principles apply in responding to test-taker characteristics that could interfere with the validity of test score interpretation. This is not to say that the response to test-taker characteristics is the same for individuals from diverse subgroups such as those defined by race, ethnicity, gender, culture, language, age, disability or socioeconomic status, but rather that these responses should be sensitive to individual characteristics that otherwise would compromise validity. Nonetheless, as discussed in the Introduction, it is important to bear in mind, when using the *Standards*, that applicability depends on context. For example, potential threats to test validity for examinees with limited English proficiency are different from those for examinees with disabilities. Moreover, threats to validity may differ even for individuals within the same subgroup. For example, individuals with diverse specific disabilities constitute the subgroup of "individuals with disabilities," and examinees classified as "limited English proficient" represent a range of language proficiency levels, educational and cultural backgrounds, and prior experiences. Further, the equivalence of the construct being assessed is a central issue in fairness, whether the context is, for example, individuals with diverse special disabilities, individuals with limited English proficiency, or individuals across countries and cultures.

As in the previous versions of the *Standards*, the current chapter addresses measurement bias as a central threat to fairness in testing. However, it also adds two major concepts that have emerged in the literature, particularly in literature regarding education, for minimizing bias and thereby increasing fairness. The first concept is *accessibility*, the notion that all test takers should have an unobstructed opportunity to demonstrate their standing on the construct(s) being measured. For example, individuals with limited English proficiency construct of interest is defined as a particular kind of language proficiency (e.g., academic language of the kind found in text books, language and vocabulary specific to workplace and employment testing).

Test Response

In some cases, construct-irrelevant variance may arise because test items elicit varieties of responses other than those intended or because items can be solved in ways that were not intended. To the extent that such responses are more typical of some subgroups than others, biased score interpretations may result. For example, some clients responding to a neuropsychological test may attempt to provide the answers they think the test administrator expects, as opposed to the answers that best describe themselves.

Construct-irrelevant components in test scores may also be associated with test response formats that pose particular difficulties or are differentially valued by particular individuals. For example, test performance may rely on some capability (e.g., English language proficiency or fine-motor coordination) that is irrelevant to the target construct(s) but nonetheless poses impediments to the test responses for some test takers not having the capability. Similarly, different values associated with the nature and degree of verbal output can influence test-taker responses. Some individuals may judge verbosity or rapid speech as rude, whereas others may regard those speech patterns as indications of high mental ability or friendliness. An individual of the first type who is evaluated with values appropriate to the second may be considered taciturn, withdrawn, or of low mental ability. Another example is a person with memory or language problems or depression; such a person's ability to communicate or show interest in communicating verbally may be constrained, which may result in interpretations of the outcomes of the assessment that are invalid and potentially harmful to the person being tested.

In the development and use of scoring rubrics, it is particularly important that credit be awarded for response characteristics central to the construct being measured and not for response characteristics that are irrelevant or tangential to the construct. Scoring rubrics may inadvertently advantage some individuals over others. For example, a scoring rubric for a constructed response item might reserve the highest score level for test takers who provide more information or elaboration than was actually requested. In this situation, test takers who simply follow instructions, or test takers who value succinctness in responses, will earn lower scores; thus, characteristics of the individuals become construct-irrelevant components of the test scores. Similarly, the scoring of open-ended responses may introduce construct-irrelevant variance for some test takers if scorers and/or automated scoring routines are not sensitive to the full diversity of ways in which individuals express their ideas. With the advent of automated scoring for complex performance tasks, for example, it is important to examine the validity of the automated scoring results for relevant subgroups in the test-taking population.

Opportunity to Learn

Finally, *opportunity to learn*—the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test—has several implications for the fair and valid interpretation of test scores for their intended uses. Individuals' prior opportunity to learn can be an important contextual factor to consider in interpreting and drawing inferences from test scores. For example, a recent immigrant who has had little prior exposure to school may not have had the opportunity to learn concepts assumed to be common knowledge by a personality inventory or ability measure, even if that measure is administered in the native language of the test taker. Similarly, as another example, there has been considerable public discussion about potential inequities in school resources available to students from traditionally disadvantaged groups, for example, racial, ethnic, language, and cultural minorities and rural students. Such inequities affect the quality of education received. To the extent that inequity exists, the validity of inferences about student ability drawn from achievement test scores may be compromised. Not taking into account prior opportunity to learn could lead to misdiagnosis, inappropriate placement, and/or inappropriate assignment of services, which could have significant consequences for an individual.

Beyond its impact on the validity of test score interpretations for intended uses, opportunity to learn has important policy and legal ramifications in education. Opportunity to learn is a fairness issue when an authority provides differential access to opportunity to learn for some individuals and then holds those individuals who have not been provided that opportunity accountable for their test performance. This problem may affect high-stakes competency tests in education, for example, when educational authorities require a certain level of test performance for high school graduation. Here, there is a fairness concern that students not be held accountable for, or face serious permanent negative consequences from, their test results when their school experiences have not provided them the opportunity to learn the subject matter covered by the test. In such cases, students' low scores may accurately reflect what they know and can do, so that, technically, the interpretation of the test results for the purpose of measuring how much the students have learned may not be biased. However, it may be considered unfair to severely penalize students for circumstances that are not under their control, that is, for not learning content that their schools have not taught. It is generally accepted that before high-stakes consequences can be imposed for failing an examination in educational settings, there must be evidence that students have been provided curriculum and instruction that incorporates the constructs addressed by the test.

[illegible]

Several important issues arise when opportunity to learn is considered as a component of fairness. First, it is difficult to define opportunity to learn in educational practice, particularly at the individual level. Opportunity is generally a matter of degree and is difficult to quantify; moreover, the measurement of some important learning outcomes may require students to work with materials that they have not seen before. Second, even if it is possible to document the topics included in the curriculum for a group of students, specific content coverage for any one student may be impossible to determine. Third, granting a diploma to a low-scoring examinee on the grounds that the student had insufficient opportunity to learn the material tested means certificating someone who has not attained the degree of proficiency the diploma is intended to signify.

It should be noted that concerns about opportunity to learn do not necessarily apply to situations where the same authority is not responsible for both the delivery of instruction and the testing and/or interpretation of results. For example, in college admissions decisions, opportunity to learn may be beyond the control of the test users and it may not influence the validity of test interpretations for their intended use (e.g., selection and/or admissions decisions). Chapter 12, "Educational Testing and Assessment," provides additional perspective on opportunity to learn.

MINISTRO EDUCAÇÃO [illegible] DECLARA

Minimizing Construct-Irrelevant Components Through Test Design and Testing Adaptations

Standardized tests should be designed to facilitate accessibility and minimize construct-irrelevant barriers for all test takers in the target population, as far as practicable. Before considering the need for any assessment adaptations for test takers who may have special needs, the assessment developer first must attempt to improve accessibility within the test itself. Some of these basic principles are included in the test design process called universal design. By using universal design, test developers begin the test development process with an eye toward maximizing fairness. Universal design emphasizes the need to develop tests that are as usable as possible for all test takers in the intended test population, regardless of characteristics such as gender, age, language background, culture, socioeconomic status, or disability.

Principles of universal design include defining constructs precisely, so that what is being measured can be clearly differentiated from test-taker characteristics that are irrelevant to the construct but that could otherwise interfere with some test