ished by
rican Educational Research Association
0 17th St., NW
hington, DC 20036

ary of Congress Card number: 99-066845
N: 0-935302-25-5

ted in the United States of America.

Standards for Educational and Psychological
ing will be under continuing review by the
ee sponsoring organizations. Comments and
gestions will be welcome and should be sent
The Committee to Develop Standards for
ucational and Psychological Testing in care of
Executive Office, American Psychological
sociation, 750 First Street, NE, Washington,
C 20002-4242.

epared by the
int Committee on Standards for Educational
d Psychological Testing of the American
ducational Research Association, the American
sychological Association, and the National
ouncil on Measurement in Education.

PREFACE v

INTRODUCTION 1
Participants in the Testing Process 2
The Purpose of the Standards 3
Categories of Standards 3
Tests and Test Uses to Which These Standards Apply 3
Cautions to be Exercised in Using the Standards 4
The Number of Standards 4
Tests as Measures of Constructs 5
Organization of This Volume 5

PART I
TEST CONSTRUCTION, EVALUATION, AND DOCUMENTATION 7

1. Validity 9
Background 9
Standards **1.1-1.24** 17

2. Reliability and Errors of Measurement 25
Background 25
Standards **2.1-2.20** 31

3. Test Development and Revision 37
Background 37
Standards **3.1-3.27** 43

4. Scales, Norms, and Score Comparability 49
Background 49
Standards **4.1-4.21** 54

5. Test Administration, Scoring, and Reporting 61
Background 61
Standards **5.1-5.16** 63

6. Supporting Documentation for Tests 67
Background 67
Standards **6.1-6.15** 68

PART II
FAIRNESS IN TESTING 71

7. Fairness in Testing and Test Use 73
Background 73
Standards **7.1-7.12** 80

8. The Rights and Responsibilities of Test Takers 85
Background 85
Standards **8.1-8.13** 86

9. Testing Individuals of Diverse Linguistic Backgrounds 91
Background 91
Standards **9.1-9.11** 97

10. Testing Individuals with Disabilities 101
Background 101
Standards **10.1-10.12** 106

PART III
TESTING APPLICATIONS 109

11. The Responsibilities of Test Users 111
Background 111
Standards **11.1-11.24** 113

12. Psychological Testing and Assessment 119
Background 119
Standards **12.1-12.20** 131

13. Educational Testing and Assessment 137
Background 137
Standards **13.1-13.19** 145

14. Testing in Employment and Credentialing 151
Background 151
Standards **14.1-14.17** 158

15. Testing in Program Evaluation and Public Policy 163
Background 163
Standards **15.1-15.13** 167

GLOSSARY 171

INDEX 185

PREFACE

There have been five earlier documents from three sponsoring organizations guiding the development and use of tests. The first of these was *Technical Recommendations for Psychological Tests and Diagnostic Techniques,* prepared by a committee of the American Psychological Association (APA) and published by that organization in 1954. The second was *Technical Recommendations for Achievement Tests,* prepared by a committee representing the American Educational Research Association (AERA) and the National Council on Measurement Used in Education (NCMUE) and published by the National Education Association in 1955. The third, which replaced the earlier two, was published by APA in 1966 and prepared by a committee representing APA, AERA, and the National Council on Measurement in Education (NCME) and called the *Standards for Educational and Psychological Tests and Manuals.* The fourth, *Standards for Educational and Psychological Tests,* was again a collaboration of AERA, APA and NCME, and was published in 1974. The fifth, *Standards for Educational and Psychological Testing,* also a joint collaboration, was published in 1985.

In 1991 APA's Committee on Psychological Tests and Assessment suggested the need to revise the 1985 *Standards.* Representatives of AERA, APA and NCME met and discussed the revision, principles that should guide that revision, and potential Joint Committee members. By 1993, the presidents of the three organizations appointed members and the Committee had its first meeting November, 1993.

The *Standards* has been developed by a joint committee appointed by AERA, APA and NCME. Members of the Committee were:

Eva Baker, *co-chair*
Paul Sackett, *co-chair*
Lloyd Bond
Leonard Feldt
David Goh
Bert Green
Edward Haertel
Jo-Ida Hansen
Sharon Johnson-Lewis
Suzanne Lane
Joseph Matarazzo
Manfred Meier
Pamela Moss
Esteban Olmedo
Diana Pullin

From 1993 to 1996 Charles Spielberger served on the Committee as co-chair. Each sponsoring organization was permitted to assign up to two liaisons to the Joint Committee's project. Liaisons served as the conduits between the sponsoring organizations and the Joint Committee. APA's liaison from its Committee on Psychological Tests and Assessments changed several times as the membership of the Committee changed.

Liaisons to the Joint Committee:

AERA -William Mehrens
APA - Bruce Bracken, Andrew Czopek, Rodney Lowman, Thomas Oakland
NCME - Daniel Eignor

APA and NCME also had committees who served to monitor the process and keep relevant parties informed.

APA Ad Hoc Committee of the Council of Representatives:

Melba Vasquez
Donald Bersoff
Stephen DeMers
James Farr
Bertram Karon
Nadine Lambert
Charles Spielberger

NCME Standards and Test Use Committee:

Gregory Cizek
Allen Doolittle
Le Ann Gamache

National Board of Medical Examiners
National Council of State Boards of Nursing

Government and Federal Agencies

Army Research Institute (ARI)
California Highway Patrol, Personnel and Training Division, Selection Research Program
City of Dallas, Civil Service Department
Commonwealth of Virginia, Department of Education
Defense Manpower Data Center (DMDC), Personnel Testing Division
Department of Defense (DOD), Office of the Assistant Secretary of Defense
Department of Education, Office of Educational Improvement, National Center for Education Statistics
Department of Justice, Immigration and Naturalization Service (INS)
Department of Labor, Employment and Training Administration (DOL/ETA)
U.S. Equal Employment Opportunity Commission (EEOC)
U.S. Office of Personnel Management (OPM), Personnel Resources & Development Center

Test Publishers/Developers

American College Testing (ACT)
CTB/McGraw-Hill
The College Board
Educational Testing Service (ETS)
Highland Publishing Company
Institute for Personality & Ability Testing (IPAT)
Professional Examination Service (PES)

Academic Institutions

Center for Creative Leadership
Gallaudet University, National Task Force on Equity in Testing Deaf Professionals
University of Haifa, Israeli Group
Kansas State University
National Center on Educational
Pennsylvania State University
University of North Carolina – Charlotte
University of Southern Mississippi, Department of Psychology

When the Joint Committee completed its task of revising the *Standards*, it then submitted its work to the three sponsoring organizations for approval. Each organization had its own governing body and mechanism for approval, as well as definitions for what their approval means.

> AERA: This endorsement carries with it the understanding that, in general, we believe the *Standards* to represent the current consensus among recognized professionals regarding expected measurement practice. Developers, sponsors, publishers, and users of tests should observe these *Standards*.
>
> APA: The APA's approval of the *Standards* means the Council adopts the document as APA policy.
>
> NCME: NCME endorses the *Standards for Educational and Psychological Testing* and recognizes that the intent of these *Standards* is to promote sound and responsible measurement practice. This endorsement carries with it a professional imperative for NCME members to attend to the *Standards*.

Although the *Standards* are prescriptive, the *Standards* itself does not contain enforcement mechanisms. These standards were formulated with the intent of being consistent with other standards, guidelines and codes of conduct published by the three sponsoring organizations, and listed below. The reader is encouraged to obtain these documents, some of which have references to testing and assessment in specific applications or settings.

The Joint Committee on the *Standards for Educational and Psychological Testing*

References

American Educational Research Association. (June, 1992). *Ethical Standards of the American Educational Research Association.* Washington, DC: Author.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. *Standards for Teacher Competence in Educational Assessment of Students.* (1990). Washington, DC: National Council on Measurement in Education.

American Psychological Association. (December, 1992). Ethical Principles of Psychologists and Code of Conduct. *American Psychologist, 47* (12), 1597-1611.

Joint Committee on Testing Practices. (1988). *Code of Fair Testing Practices in Education.* Washington, DC: American Psychological Association.

National Council on Measurement in Education. (1995). *Code of Professional Responsibilities in Educational Measurement.* Washington, DC: Author.

Educational and psychological testing and assessment are among the most important contributions of behavioral science to our society, providing fundamental and significant improvements over previous practices. Although not all tests are well-developed nor are all testing practices wise and beneficial, there is extensive evidence documenting the effectiveness of well-constructed tests for uses supported by validity evidence. The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education and employment. The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions. The intent of the *Standards* is to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices.

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, teachers, educational administrators, job applicants, employees, clients, patients, supervisors, executives, and evaluators, among others. The institutions affected include schools, colleges, businesses, industry, 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.

The interests of the various parties involved in the testing process are usually, but not always, 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.

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 use the test results for some decision-making purpose; (e) those who interpret test results for clients; (f) those who take the test by choice, direction, or necessity; (g) those who sponsor tests, which may be 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.

These roles are sometimes combined and sometimes further divided. 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 industrial 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, advertised, and distributed by an independent publisher, though the publisher may play an active role in the test development. Given this intermingling of roles, it is difficult to assign precise responsibility for addressing various standards to specific participants in the testing process.

This document begins with a series of chapters on the test development process, which focus primarily on the responsibilities of test developers, and then turns to chapters

on specific uses and applications, which focus primarily on responsibilities of test users. One chapter is devoted specifically to the rights and responsibilities of test takers.

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

The Purpose of the Standards

The purpose of publishing the *Standards* is to provide criteria for the evaluation of tests, testing practices, and the effects of test use. Although the evaluation of the appropriateness of a test or testing application should depend heavily on professional judgment, the *Standards* provides a frame of reference to assure that relevant issues are addressed. It is hoped that all professional test developers, sponsors, publishers, and users will adopt the *Standards* and encourage others to do so.

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 debate may be fully informed.

Categories of Standards

The 1985 *Standards* designated each standard as "primary" (to be met by all tests before operational use), "secondary" (desirable, but not feasible in certain situations), or "conditional" (importance varies with application). The present *Standards* continues the tradition of expecting test developers and users to consider all standards before operational use; however, the *Standards* does not continue the practice of designating levels of importance. Instead, the text of each standard, and any accompanying commentary, discusses the conditions under which a standard is relevant. It was not the case that under the 1985 *Standards* test developers and users were obligated to attend only to the primary standards. Rather, the term "conditional" meant that a standard was primary in some settings and secondary in others, thus requiring careful consideration of the applicability of each standard for a given setting.

The absence of designations such as "primary" or "conditional" should not be taken to imply that all standards are equally significant in any given situation. 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 some standards that need to be addressed in particular situations.

Further, the current *Standards* does not include standards considered secondary or "desirable." The continued use of the secondary designation would risk encouraging both the expansion of the *Standards* to encompass large numbers of "desirable" standards and the inappropriate assumption that any guideline not included in the *Standards* as at least "secondary" was inconsequential.

Unless otherwise specified in the standard or commentary, and with the caveats outlined below, standards should be met before operational test use. This means that 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 why adherence to the standard is unnecessary. It is also possible that there may be occasions when technical feasibility may influence whether a standard can be met prior to operational test use. For example, some standards may call for analyses of data that may not be available at the point of initial operational test use. If test developers, users, and, when applicable, sponsors have deemed a standard to be inapplicable or unfeasible, they should be able, if called upon, to explain the basis for their decision. However, there is no expectation that documentation be routinely available of the decisions related to each standard.

Tests and Test Uses to Which These Standards Apply

A test is an evaluative 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. While the label *test* is ordinarily reserved for instruments on which responses are evaluated for their correctness or quality and the terms *scale* or *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 *test* and *assessment*. *Assessment* is a broader term, commonly referring to a process that integrates test information with information from other sources (e.g., information from the individual's social, educational, employment, or psychological history). The applicability of the *Standards* to an evaluation device or method is not altered by the label applied to it (e.g., test, assessment, scale, inventory).

Tests differ on a number of dimensions: the mode in which test materials are presented (paper and pencil, oral, computerized administration, and so on); 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 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-taker responses to test materials are evaluated and scored. As noted in prior versions of the *Standards*, the same general types of information are needed for all varieties of tests.

The precise demarcation between those 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, it may also be usefully applied in varying degrees to a broad range of less formal assessment techniques. Admittedly, it will generally not be possible to apply the *Standards* rigorously to unstandardized questionnaires or to the broad range of unstructured behavior samples used in some forms of clinic- and school-based psychological assessment (e.g., an intake interview), and to instructor-made tests that are used to evaluate student performance in education and training. It is useful to distinguish between devices that lay claim to the concepts and techniques of the field of educational and psychological testing from those which represent nonstandardized 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 evalu-

ates the performance of subordinates, 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*.

Cautions to be Exercised in Using the Standards

Several cautions are important to avoid misinterpreting the *Standards*:

1) Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability 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 involves (a) professional judgment that is based on a knowledge of behavioral science, psychometrics, and the community standards in the professional field to which the tests apply; (b) the degree to which the intent of the standard has been satisfied by the test developer and user; (c) the alternatives that are readily available; and (d) research and experiential evidence regarding feasibility of meeting the standard.

2) When tests are at issue in legal proceedings and other venues 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.

3) Claims by test developers or test users that a test, manual, or procedure satisfies or follows these standards 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.

4) These standards are concerned with a field that is evolving. Consequently, there is a continuing need to monitor changes in the field and to revise this document as knowledge develops.

5) Prescription of 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" always should be understood.

The standards do not attempt to repeat or to incorporate the many legal or regulatory requirements that might be relevant to the issues they address. 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. Where it is apparent that one or more standards or comments 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 that no relevant requirement exists. In all situations, participants in the testing process should separately consider and, where appropriate, obtain legal advice on legal and regulatory requirements.

The Number of Standards

The number of standards has increased from the 1985 *Standards* for a variety of reasons. First, and most importantly, new developments have led to the addition of new standards. Commonly these deal with new types of tests or new uses for existing tests, rather than being broad standards applicable to all tests. Second, on the basis of recognition that some users of the *Standards* may turn only to chapters directly relevant to a given application, certain standards are repeated in different chapters. When such repetition occurs, the essence of the standard is the same. Only the wording, area of application, or elaboration in the comment is changed. Third, standards dealing with important nontechnical issues, such as avoiding conflicts of interest and equitable treatment of all test takers, have been added. Although such topics have not been addressed in prior versions of the *Standards*, they are not likely to be viewed as imposing burdensome new requirements. Thus the increase in the number of standards does not per se signal an increase in the obligations placed on test developers and test users.

Tests as Measures of Constructs

We depart from some historical uses of the term "construct," which reserve the term for characteristics that are not directly observable, but which are inferred from interrelated sets of observations. This historical perspective invites confusion. Some tests are viewed as measures of constructs, while others are not. In addition, considerable debate has ensued as to whether certain characteristics measured by tests are properly viewed as constructs. Furthermore, the types of validity evidence thought to be suitable can differ as a result of whether a given test is viewed as measuring a construct.

We use the term *construct* more broadly as the concept or characteristic that a test is designed to measure. Rarely, if ever, is there a single possible meaning that can be attached to a test score or a pattern of test responses. Thus, it is always incumbent on a testing professional to specify the construct interpretation that will be made on the basis of the score or response pattern. The notion that some tests are not under the purview of the *Standards* because they do not measure constructs is contrary to this use of the term. Also, as detailed in chapter 1, evolving conceptualizations of the concept of validity no longer speak of different types of validity but speak instead of different lines of validity evidence, all in service of providing information relevant to a specific intended interpretation of test scores. Thus, many lines of evidence can contribute to an understanding of the construct meaning of test scores.

Organization of This Volume

Part I of the *Standards*, "Test Construction, Evaluation, and Documentation," contains standards for validity (ch. 1); reliability and errors of measurement (ch. 2); test development and revision (ch. 3); scaling, norming, and score comparability (ch. 4); test administration, scoring, and reporting (ch. 5); and supporting documentation for tests (ch. 6). Part II addresses "Fairness in Testing," and contains standards on fairness and bias (ch. 7); the rights and responsibilities of test takers (ch. 8); testing individuals of diverse linguistic backgrounds (ch. 9); and testing individuals with disabilities (ch. 10). Part III treats specific "Testing Applications," and contains standards involving general responsibilities of test users (ch. 11); psychological testing and assessment (ch. 12); educational testing and assessment (ch. 13); testing in employment and credentialing (ch. 14); and testing in program evaluation and public policy (ch. 15).

Each chapter begins with introductory text that provides background for the standards that follow. This revision of the *Standards* contains more extensive introductory text material than its predecessor. Recognizing the common use of the *Standards* in the education of future test developers and users, the committee opted to provide a context for the standards themselves by pre-

senting more background material than in previous versions. This text is designed to assist in the interpretation of the standards that follow in each chapter. Although the text is at times prescriptive and exhortatory, it should not be interpreted as imposing additional standards.

The *Standards* also contains an index and includes a glossary that provides definitions for terms as they are specifically used in this volume.

PART I

Test Construction, Evaluation, and Documentation

Background

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated.

Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the construct or concepts the test is intended to measure. Examples of constructs are mathematics achievement, performance as a computer technician, depression, and self-esteem. To support test development, the proposed interpretation is elaborated by describing its scope and extent and by delineating the aspects of the construct that are to be represented. The detailed description provides a conceptual framework for the test, delineating the knowledge, skills, abilities, processes, or characteristics to be assessed. The framework indicates how this representation of the construct is to be distinguished from other constructs and how it should relate to other variables.

The conceptual framework is partially shaped by the ways in which test scores will be used. For instance, a test of mathematics achievement might be used to place a student in an appropriate program of instruction, to endorse a high school diploma, or to inform a college admissions decision. Each of these uses implies a somewhat different interpretation of the mathematics achievement test scores: that a student will benefit from a particular instructional intervention, that a student has mastered a specified curriculum, or that a student is likely to be successful with college-level work. Similarly, a test of self-esteem might be used for psychological counseling, to inform a decision about employment, or for the basic scientific purpose of elaborating the construct of self-esteem. Each of these potential uses shapes the specified framework and the proposed interpretation of the test's scores and also has implications for test development and evaluation.

Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. The conceptual framework points to the kinds of evidence that might be collected to evaluate the proposed interpretation in light of the purposes of testing. As validation proceeds, and new evidence about the meaning of a test's scores becomes available, revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test.

The wide variety of tests and circumstances makes it natural that some types of evidence will be especially critical in a given case, whereas other types will be less useful. The decision about what types of evidence are important for validation in each instance can be clarified by developing a set of propositions that support the proposed interpretation for the particular purpose of testing. For instance, when a mathematics achievement test is used to assess readiness for an advanced course, evidence for the following propositions might be deemed necessary: (a) that certain skills are prerequisite for the advanced course; (b) that the content domain of the test is consistent with these prerequisite skills; (c) that test scores can be generalized across relevant sets of items; (d) that test scores are not unduly influenced by ancillary variables,

advanced course can be validly assessed; and (f) that examinees with high scores on the test will be more successful in the advanced course than examinees with low scores on the test. Examples of propositions in other testing contexts might include, for instance, the proposition that examinees with high general anxiety scores experience significant anxiety in a range of settings, the proposition that a child's score on an intelligence scale is strongly related to the child's academic performance, or the proposition that a certain pattern of scores on a neuropsychological battery indicates impairment characteristic of brain injury. The validation process evolves as these propositions are articulated and evidence is gathered to evaluate their soundness.

Identifying the propositions implied by a proposed test interpretation can be facilitated by considering rival hypotheses that may challenge the proposed interpretation. It is also useful to consider the perspectives of different interested parties, existing experience with similar tests and contexts, and the expected consequences of the proposed test use. Plausible rival hypotheses can often be generated by considering whether a test measures less or more than its proposed construct. Such concerns are referred to as *construct underrepresentation* and *construct-irrelevant* variance.

Construct underrepresentation refers to the degree to which a test fails to capture important aspects of the construct. It implies a narrowed meaning of test scores because the test does not adequately sample some types of content, engage some psychological processes, or elicit some ways of responding that are encompassed by the intended construct. Take, for example, a test of reading comprehension intended to measure children's ability to read and interpret stories with understanding. A particular test might underrepresent the intended construct because it did not contain a sufficient variety of reading material. As another example, a test of anxiety might measure only physiological reactions and not emotional, cognitive, or situational components.

Construct-irrelevant variance refers to the degree to which test scores are affected by processes that are extraneous to its intended construct. The test scores may be systematically influenced to some extent by components that are not part of the construct. In the case of a reading comprehension test, construct-irrelevant components might include an emotional reaction to the test content, familiarity with the subject matter of the reading passages on the test, or the writing skill needed to compose a response. Depending on the detailed definition of the construct, vocabulary knowledge or reading speed might also be irrelevant components. On a test of anxiety, a response bias to underreport anxiety might be considered a source of construct-irrelevant variance.

Nearly all tests leave out elements that some potential users believe should be measured and include some elements that some potential users consider inappropriate. Validation involves careful attention to possible distortions in meaning arising from inadequate representation of the construct and also to aspects of measurement such as test format, administration conditions, or language level that may materially limit or qualify the interpretation of test scores. That is, the process of validation may lead to revisions in the test, the conceptual framework of the test, or both. The revised test would then need validation.

When propositions have been identified that would support the proposed interpretation of test scores, validation can proceed by developing empirical evidence, examining relevant literature, and/or conducting logical analyses to evaluate each of these propositions. Empirical evidence may include both local evidence, produced within the contexts where the test will be used, and evidence from similar testing

evidence from similar tests and contexts can enhance the quality of the validity argument, especially when current data are limited.

Because a validity argument typically depends on more than one proposition, strong evidence in support of one in no way diminishes the need for evidence to support others. For example, a strong predictor-criterion relationship in an employment setting is not sufficient to justify test use for selection without considering the appropriateness and meaningfulness of the criterion measure. Professional judgment guides decisions regarding the specific forms of evidence that can best support the intended interpretation and use. As in all scientific endeavors, the quality of the evidence is primary. A few lines of solid evidence regarding a particular proposition are better than numerous lines of evidence of questionable quality.

Validation is the joint responsibility of the test developer and the test user. The test developer is responsible for furnishing relevant evidence and a rationale in support of the intended test use. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used. When the use of a test differs from that supported by the test developer, the test user bears special responsibility for validation. The standards apply to the validation process, for which the appropriate parties share responsibility. It should be noted that important contributions to the validity evidence are made as other researchers report findings of investigations that are related to the meaning of scores on the test.

Sources of Validity Evidence

The following sections outline various sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes. These sources of evidence may illuminate different aspects of validity,

but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose. Like the 1985 *Standards*, this edition refers to types of validity evidence, rather than distinct types of validity. To emphasize this distinction, the treatment that follows does not follow traditional nomenclature (i.e., the use of the terms *content validity* or *predictive validity*). The glossary contains definitions of the traditional terms, explicating the difference between traditional and current use.

Evidence Based on Test Content

Important validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure. Test content refers to the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring. Test developers often work from a specification of the content domain. The content specification carefully describes the content in detail, often with a classification of areas of content and types of items. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on content can also come from expert judgments of the relationship between parts of the test and the construct. For example, in developing a licensure test, the major facets of the specific occupation can be specified, and experts in that occupation can be asked to assign test items to the categories defined by those facets. They, or other qualified experts, can then judge the representativeness of the chosen set of items. Sometimes rules or algorithms can be constructed to select or generate items that differ systematically on the various facets of content, according to specifications.

Some tests are based on systematic observations of behavior. For example, a listing of the tasks comprising a job domain may be developed from observations of behavior in a job, together with judgments of subject-matter experts. Expert judgments can be used to assess the relative importance, criticality, and/or frequency of the various tasks. A job sample test can then be constructed from a random or stratified sampling of tasks rated highly on these characteristics. The test can then be administered under standardized conditions in an off-the-job setting.

The appropriateness of a given content domain is related to the specific inferences to be made from test scores. Thus, when considering an available test for a purpose other than that for which it was first developed, it is especially important to evaluate the appropriateness of the original content domain for the proposed new use. In educational program evaluations, for example, tests may properly cover material that receives little or no attention in the curriculum, as well as that toward which instruction is directed. Policymakers can then evaluate student achievement with respect to both content neglected and content addressed. On the other hand, when student mastery of a delivered curriculum is tested for purposes of informing decisions about individual students, such as promotion or graduation, the framework elaborating a content domain is appropriately limited to what students have had an opportunity to learn from the curriculum as delivered.

Evidence about content can be used, in part, to address questions about differences in the meaning or interpretation of test scores across relevant subgroups of examinees. Of particular concern is the extent to which construct underrepresentation or construct-irrelevant components may give an unfair advantage or disadvantage to one or more subgroups of examinees. Careful review of the construct and test content domain by a diverse panel of experts may point to potential sources of irrelevant difficulty (or easiness) that require further investigation.

Evidence Based on Response Processes

Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees. For instance, if a test is intended to assess mathematical reasoning, it becomes important to determine whether examinees are, in fact, reasoning about the material given instead of following a standard algorithm. For another instance, scores on a scale intended to assess the degree of an individual's extroversion or introversion should not be strongly influenced by social conformity.

Evidence based on response processes generally comes from analyses of individual responses. Questioning test takers about their performance strategies or responses to particular items can yield evidence that enriches the definition of a construct. Maintaining records that monitor the development of a response to a writing task, through successive written drafts or electronically monitored revisions, for instance, also provides evidence of process. Documentation of other aspects of performance, like eye movements or response times, may also be relevant to some constructs. Inferences about processes involved in performance can also be developed by analyzing the relationship among parts of the test and between the test and other variables. Wide individual differences in process can be revealing and may lead to reconsideration of certain test formats.

Evidence of response processes can contribute to questions about differences in meaning or interpretation of test scores across relevant subgroups of examinees. Process studies involving examinees from different subgroups can assist in determining the extent to which capabilities irrelevant or ancillary to the construct may be differentially influencing their performance.

Studies of response processes are not limited to the examinee. Assessments often rely on observers or judges to record and/or evaluate examinees' performances or products. In such cases, relevant validity evidence includes the extent to which the processes of observers or judges are consistent with the intended interpretation of scores. For instance, if judges are expected to apply particular criteria in scoring examinees' performances, it is important to ascertain whether they are, in fact, applying the appropriate criteria and not being influenced by factors that are irrelevant to the intended interpretation. Thus, validation may include empirical studies of how observers or judges record and evaluate data along with analyses of the appropriateness of these processes to the intended interpretation or construct definition.

Evidence Based on Internal Structure

Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based. The conceptual framework for a test may imply a single dimension of behavior, or it may posit several components that are each expected to be homogeneous, but that are also distinct from each other. For example, a measure of discomfort on a health survey might assess both physical and emotional health. The extent to which item interrelationships bear out the presumptions of the framework would be relevant to validity.

The specific types of analysis and their interpretation depend on how the test will be used. For example, if a particular application posited a series of test components of increasing difficulty, empirical evidence of the extent to which response patterns conformed to this expectation would be provided. A theory that posited unidimensionality would call for evidence of item homogeneity. In this case, the item interrelationships also provide an estimate of score reliability, but such an index would be inappropriate for tests with a more complex internal structure.

Some studies of the internal structure of tests are designed to show whether particular items may function differently for identifiable subgroups of examinees. Differential item functioning occurs when different groups of examinees with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item. This issue is discussed in chapters 3 and 7. However, differential item functioning is not always a flaw or weakness. Subsets of items that have a specific characteristic in common (e.g., specific content, task representation) may function differently for different groups of similarly scoring examinees. This indicates a kind of multidimensionality that may be unexpected or may conform to the test framework.

Evidence Based on Relations to Other Variables

Analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence. External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs. Measures other than test scores, such as performance criteria, are often used in employment settings. Categorical variables, including group membership variables, become relevant when the theory underlying a proposed test use suggests that group differences should be present or absent if a proposed test interpretation is to be supported. Evidence based on relationships with other variables addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretations.

Convergent and discriminant evidence. Relationships between test scores and other measures intended to assess similar constructs provide convergent evidence, whereas relationships between test scores and measures purportedly of different constructs provide discriminant evidence. For instance, within some theoretical frameworks, scores on a multiple-choice test of reading comprehension might be expected to relate closely (convergent evidence) to other measures of reading comprehension based on other methods, such as essay responses; conversely, test scores might be expected to relate less closely (discriminant evidence) to measures of other skills, such as logical reasoning. Relationships among different methods of measuring the construct can be especially helpful in sharpening and elaborating score meaning and interpretation.

Evidence of relations with other variables can involve experimental as well as correlational evidence. Studies might be designed, for instance, to investigate whether scores on a measure of anxiety improve as a result of some psychological treatment or whether scores on a test of academic achievement differentiate between instructed and noninstructed groups. If performance increases due to short-term coaching are viewed as a threat to validity, it would be useful to investigate whether coached and uncoached groups perform differently.

Test-criterion relationships. Evidence of the relation of test scores to a relevant criterion may be expressed in various ways, but the fundamental question is always: How accurately do test scores predict criterion performance? The degree of accuracy deemed necessary depends on the purpose for which the test is used.

The criterion variable is a measure of some attribute or outcome that is of primary interest, as determined by test users, who may be administrators in a school system, the management of a firm, or clients. The choice of the criterion and the measurement procedures used to obtain criterion scores are of central importance. The value of a test-criterion study depends on the relevance, reliability, and validity of the interpretation based on the criterion measure for a given testing application.

Historically, two designs, often called predictive and concurrent, have been distinguished for evaluating test-criterion relationships. A predictive study indicates how accurately test data can predict criterion scores that are obtained at a later time. A concurrent study obtains predictor and criterion information at about the same time. When prediction is actually contemplated, as in education or employment settings, or in planning rehabilitation regimens, predictive studies can retain the temporal differences and other characteristics of the practical situation. Concurrent evidence, which avoids temporal changes, is particularly useful for psychodiagnostic tests or to investigate alternative measures of some specified construct. In general, the choice of research strategy is guided by prior evidence of the extent to which predictive and concurrent studies yield the same or different results in the domain.

Test scores are sometimes used in allocating individuals to different treatments, such as different jobs within an institution, in a way that is advantageous for the institution and for the individuals. In that context, evidence is needed to judge the suitability of using a test when classifying or assigning a person to one job versus another or to one treatment versus another. Classification decisions are supported by evidence that the relationship of test scores to performance criteria is different for different treatments. It is possible for tests to be highly predictive of performance for different education programs or jobs without providing the information necessary to make a comparative judgment of the efficacy of assignments or treatments. In general, decision rules for selection or placement are also influenced by the number of persons to be accepted or the numbers that can be accommodated in alternative placement categories.

Evidence about relations to other variables is also used to investigate questions of differential prediction for groups. For instance, a finding that the relation of test scores to a relevant criterion variable differs from one group to another may imply that the meaning of the scores is not the same for members of the different groups, perhaps due to construct underrepresentation or construct-irrelevant components. However, the difference may also imply that the criterion has different meaning for different groups. The differences in test-criterion relationships can also arise from measurement error, especially when group means differ, so such differences do not necessarily indicate differences in score meaning. (See chapter 7.)

Validity generalization. An important issue in educational and employment settings is the degree to which evidence of validity based on test-criterion relations can be generalized to a new situation without further study of validity in that new situation. When a test is used to predict the same or similar criteria (e.g., performance of a given job) at different times or in different places, it is typically found that observed test-criterion correlations vary substantially. In the past, this has been taken to imply that local validation studies are always required. More recently, meta-analytic analyses have shown that in some domains, much of this variability may be due to statistical artifacts such as sampling fluctuations and variations across validation studies in the ranges of test scores and in the reliability of criterion measures. When these and other influences are taken into account, it may be found that the remaining variability in validity coefficients is relatively small. Thus, statistical summaries of past validation studies in similar situations may be useful in estimating test-criterion relationships in a new situation. This practice is referred to as the study of validity generalization.

In some circumstances, there is a strong basis for using validity generalization. This would be the case where the meta-analytic database is large, where the meta-analytic data adequately represent the type of situation to which one wishes to generalize, and where correction for statistical artifacts produces a clear and consistent pattern of validity evidence. In such circumstances, the informational value of a local validity study may be relatively limited. In other circumstances, the inferential leap required for generalization may be much larger. The meta-analytic database may be small, the findings may be less consistent, or the new situation may involve features markedly different from those represented in the meta-analytic database. In such circumstances, situation-specific evidence of validity will be relatively more informative. Although research on validity generalization shows that results of a single local validation study may be quite imprecise, there are situations where a single study, carefully done, with adequate sample size, provides sufficient evidence to support test use in a new situation. This highlights the importance of examining carefully the comparative informational value of local versus meta-analytic studies.

In conducting studies of the generalizability of validity evidence, the prior studies that are included may vary according to several situational facets. Some of the major facets are (a) differences in the way the predictor construct is measured, (b) the type of job or curriculum involved, (c) the type of criterion measure used, (d) the type of test takers, and (e) the time period in which the study was conducted. In any particular study of validity generalization, any number of these facets might vary, and a major objective of the study is to determine empirically the extent to which variation in these facets affects the test-criterion correlations obtained.

The extent to which predictive or concurrent evidence of validity generalization can

be used in new situations is in large measure a function of accumulated research. Although evidence of generalization can often help to support a claim of validity in a new situation, the extent of available data limits the extent to which the claim can be sustained.

The above discussion focuses on the use of cumulative databases to estimate predictor-criterion relationships. Meta-analytic techniques can also be used to summarize other forms of data relevant to other inferences one may wish to draw from test scores in a particular application, such as effects of coaching and effects of certain alterations in testing conditions to accommodate test takers with certain disabilities.

Evidence Based on Consequences of Testing

An issue receiving attention in recent years is the incorporation of the intended and unintended consequences of test use into the concept of validity. Evidence about consequences can inform validity decisions. Here, however, it is important to distinguish between evidence that is directly relevant to validity and evidence that may inform decisions about social policy but falls outside the realm of validity.

Distinguishing between issues of validity and issues of social policy becomes particularly important in cases where differential consequences of test use are observed for different identifiable groups. For example, concerns have been raised about the effect of group differences in test scores on employment selection and promotion, the placement of children in special education classes, and the narrowing of a school's curriculum to exclude learning of objectives that are not assessed. Although information about the consequences of testing may influence decisions about test use, such consequences do not in and of themselves detract from the validity of intended test interpretations. Rather, judgments of validity or invalidity in the light of testing consequences depend on a more searching inquiry into the sources of those consequences.

Take, as an example, a finding of different hiring rates for members of different groups as a consequence of using an employment test. If the difference is due solely to an unequal distribution of the skills the test purports to measure, and if those skills are, in fact, important contributors to job performance, then the finding of group differences per se does not imply any lack of validity for the intended inference. If, however, the test measured skill differences unrelated to job performance (e.g., a sophisticated reading test for a job that required only minimal functional literacy), or if the differences were due to the test's sensitivity to some examinee characteristic not intended to be part of the test construct, then validity would be called into question, even if test scores correlated positively with some measure of job performance. Thus, evidence about consequences may be directly relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components. Evidence about consequences that cannot be so traced—that in fact reflects valid differences in performance—is crucial in informing policy decisions but falls outside the technical purview of validity.

Tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores. A few of the many possible benefits are selection of efficacious treatments for therapy, placement of workers in suitable jobs, prevention of unqualified individuals from entering a profession, or improvement of classroom instructional practices. A fundamental purpose of validation is to indicate whether these specific benefits are likely to be realized. Thus, in the case of a test used in placement decisions, the validation would be informed by evidence that alternative placements, in fact, are differentially beneficial to the persons and the institution. In the case of employment testing, if a test publisher claims that use of the test will result in reduced employee training costs, improved workforce efficiency, or some other benefit, then the validation would be informed by evidence in support of that claim.

Claims are sometimes made for benefits of testing that go beyond direct uses of the test scores themselves. Educational tests, for example, may be advocated on the grounds that their use will improve student motivation or encourage changes in classroom instructional practices by holding educators accountable for valued learning outcomes. Where such claims are central to the rationale advanced for testing, the direct examination of testing consequences necessarily assumes even greater importance. The validation process in such cases would be informed by evidence that the anticipated benefits of testing are being realized.

Integrating the Validity Evidence

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. It encompasses evidence gathered from new studies and evidence available from earlier reported research. The validity argument may indicate the need for refining the definition of the construct, may suggest revisions in the test or other aspects of the testing process, and may indicate areas needing further study.

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees, as described in subsequent chapters of the *Standards*.

Standard 1.1

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

Comment: The rationale should indicate what propositions are necessary to investigate the intended interpretation. The comprehensive summary should combine logical analysis with empirical evidence to provide support for the test rationale. Evidence may come from studies conducted locally, in the setting where the test is to be used; from specific prior studies; or from comprehensive statistical syntheses of available studies meeting clearly specified criteria. No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test use determine the value of a particular kind of evidence. A presentation of empirical evidence on any point should give due weight to all relevant findings in the scientific literature, including those inconsistent with the intended interpretation or use. Test developers have the responsibility to provide support for their own recommendations, but test users are responsible for evaluating the quality of the validity evidence provided and its relevance to the local situation.

Standard 1.2

The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described.

Comment: Statements about validity should refer to particular interpretations and uses. It is incorrect to use the unqualified phrase "the validity of the test." No test is valid for all purposes or in all situations. Each recom-

mended use or interpretation requires validation and should specify in clear language the population for which the test is intended, the construct it is intended to measure, and the manner and contexts in which test scores are to be employed.

Standard 1.3

If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.

Comment: If past experience suggests that a test is likely to be used inappropriately for certain kinds of decisions, specific warnings against such uses should be given. On the other hand, no two situations are ever identical, so some generalization by the user is always necessary. Professional judgment is required to evaluate the extent to which existing validity evidence supports a given test use.

Standard 1.4

If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary.

Comment: Professional judgment is required to evaluate the extent to which existing validity evidence applies in the new situation and to determine what new evidence may be needed. The amount and kinds of new evidence required may be influenced by experience with similar prior test uses or interpretations and by the amount, quality, and relevance of existing data.

Standard 1.5

The composition of any sample of examinees from which validity evidence is obtained should be described in as much detail as is practical, including major relevant sociodemographic and developmental characteristics.

Comment: Statistical findings can be influenced by factors affecting the sample on which the results are based. When the sample is intended to represent a population, that population should be described, and attention should be drawn to any systematic factors that may limit the representativeness of the sample. Factors that might reasonably be expected to affect the results include self-selection, attrition, linguistic prowess, disability status, and exclusion criteria, and others. If the subjects of a validity study are patients, for example, then the diagnoses of the patients are important, as well as other characteristics, such as the severity of the diagnosed condition. For tests used in industry, the employment status (e.g., applicants versus current job holders), the general level of experience and educational background and the gender and ethnic composition of the sample may be relevant information. For tests used in educational settings, relevant information may include educational background, developmental level, community characteristics, or school admissions policies, as well as the gender and ethnic composition of the sample. Sometimes restrictions about privacy preclude obtaining such population information.

Standard 1.6

When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

Comment: For example, test developers might provide a logical structure that maps the items on the test to the content domain, illustrating the relevance of each item and the adequacy with which the set of items represents the content domain. Areas of the content domain that are not included among the test items could be indicated as well.

Standard 1.7

When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

Comment: Systematic collection of judgments or opinions may occur at many points in test construction (e.g., in eliciting expert judgments of content appropriateness or adequate content representation), in formulating rules or standards for score interpretation (e.g., in setting cut scores), or in test scoring (e.g., rating of essay responses). Whenever such procedures are employed, the quality of the resulting judgments is important to the validation. It may be entirely appropriate to have experts work together to reach consensus, but it would not then be appropriate to treat their respective judgments as statistically independent.

Standard 1.8

If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.

Comment: If the test specification delineates the processes to be assessed, then evidence is needed that the test items do, in fact, tap the intended processes.

Standard 1.9

If a test is claimed to be essentially unaffected by practice and coaching, then the sensitivity of test performance to change with these forms of instruction should be documented.

Comment: Materials to aid in score interpretation should summarize evidence indicating the degree to which improvement with practice or coaching can be expected. Also, materials written for test takers should provide practical guidance about the value of test preparation activities, including coaching.

Standard 1.10

When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretation should be provided. When interpretation of individual item responses is likely but is not recommended by the developer, the user should be warned against making such interpretations.

Comment: Users should be given sufficient guidance to enable them to judge the degree of confidence warranted for any use or interpretation recommended by the test developer. Test manuals and score reports should discourage overinterpretation of information that may be subject to considerable error. This is especially important if interpretation

of performance on isolated items, small subsets of items, or subtest scores is suggested.

Standard 1.11

If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided.

Comment: It might be claimed, for example, that a test is essentially unidimensional. Such a claim could be supported by a multivariate statistical analysis, such as a factor analysis, showing that the score variability attributable to one major dimension was much greater than the score variability attributable to any other identified dimension. When a test provides more than one score, the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed.

Standard 1.12

When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

Comment: When a test provides more than one score, the distinctiveness of the separate scores should be demonstrated, and the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed. Moreover, evidence for the validity of interpretations of two separate scores would not necessarily justify an interpretation of the difference between them. Rather, the rationale and supporting evidence must pertain directly to the specific score or score combination to be interpreted or used.

Standard 1.13

When validity evidence includes statistical analyses of test results, either alone or together with data on other variables, the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions. Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance.

Comment: Such conditions might include (but would not be limited to) the following: examinee motivation or prior preparation, the distribution of test scores over examinees, the time allowed for examinees to respond or other administrative conditions, examiner training or other examiner characteristics, the time intervals separating collection of data on different measures, or conditions that may have changed since the validity evidence was obtained.

Standard 1.14

When validity evidence includes empirical analyses of test responses together with data on other variables, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent.

Comment: The patterns of association between and among scores on the instrument under study and other variables should be consistent with theoretical expectations. The additional variables might be demographic characteristics, indicators of treatment conditions, or scores on other measures. They might include intended measures of the same construct or of different constructs. The reliability of scores from such other measures and the validity of intended interpretations of scores from these measures are an important part of the validity evidence for the instrument under study. If such variables include composite scores, the construction of the composites should be explained. In addition to considering the properties of each variable in isolation, it is important to guard against faulty interpretations arising from spurious sources of dependency among measures, including correlated errors or shared variance due to common methods of measurement or common elements.

Standard 1.15

When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.

Comment: Regression equations are more useful than correlation coefficients, which are generally insufficient to fully describe patterns of association between tests and other variables. Means, standard deviations, and other statistical summaries are needed, as well as information about the distribution of criterion performances conditional upon a given test score. Evidence of overall association between variables should be supplemented by information about the form of that association and about the variability associated with that association in different ranges of test scores. Note that data collections employing examinees selected for their extreme scores on one or more measures (extreme groups) typically cannot provide adequate information about the association.

Standard 1.16

When validation relies on evidence that test scores are related to one or more criterion variables, information about the suitability and technical quality of the criteria should be reported.

Comment: The description of each criterion variable should include evidence concerning its reliability, the extent to which it represents the intended construct (e.g., job performance), and the extent to which it is likely to be influenced by extraneous sources of variance. Special attention should be given to sources that previous research suggests may introduce extraneous variance that might bias the criterion for or against identifiable groups.

Standard 1.17

If test scores are used in conjunction with other quantifiable variables to predict some outcome or criterion, regression (or equivalent) analyses should include those additional relevant variables along with the test scores.

Comment: In general, if several predictors of some criterion are available, the optimum combination of predictors cannot be determined solely from separate, pairwise examinations of the criterion variable with each separate predictor in turn. It is often informative to estimate the increment in predictive accuracy that may be expected when each variable, including the test score, is introduced in addition to all other available variables. Analyses involving multiple predictors should be verified by cross-validation or equivalent analysis whenever feasible, and the precision of estimated regression coefficients should be reported.

Standard 1.18

When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coeffi-

cients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported.

Comment: The correlation between two variables, such as test scores and criterion measures, depends on the range of values on each variable. For example, the test scores and the criterion values of selected applicants will typically have a smaller range than the scores of all applicants. Statistical methods are available for adjusting the correlation to reflect the population of interest rather than the sample available. Such adjustments are often appropriate, as when comparing results across various situations. Reporting an adjusted correlation should be accompanied by a statement of the method and the statistics used in making the adjustment.

Standard 1.19

If a test is recommended for use in assigning persons to alternative treatments or is likely to be so used, and if outcomes from those treatments can reasonably be compared on a common criterion, then, whenever feasible, supporting evidence of differential outcomes should be provided.

Comment: If a test is used for classification into alternative occupational, therapeutic, or educational programs, it is not sufficient just to show that the test predicts treatment outcomes. Support for the validity of the classification procedure is provided by showing that the test is useful in determining which persons are likely to profit differentially from one treatment or another. Treatment categories may have to be combined to assemble sufficient cases for statistical analysis. It is recognized, however, that such research may not be feasible, because ethical and legal constraints on differential assignments may forbid control groups.

Standard 1.20

When a meta-analysis is used as evidence of the strength of a test-criterion relationship, the test and the criterion variables in the local situation should be comparable with those in the studies summarized. If relevant research includes credible evidence that any other features of the testing application may influence the strength of the test-criterion relationship, the correspondence between those features in the local situation and in the meta-analysis should be reported. Any significant disparities that might limit the applicability of the meta-analytic findings to the local situation should be noted explicitly.

Comment: The meta-analysis should incorporate all available studies meeting explicitly stated inclusion criteria. Meta-analytic evidence used in test validation typically is based on a number of tests measuring the same or very similar constructs and criterion measures that likewise measure the same or similar constructs. A meta-analytic study may also be limited to a single test and a single criterion. For each study included in the analysis, the test-criterion relationship is expressed in some common metric, often as an *effect size*. The strength of the test-criterion relationship may be moderated by features of the situation in which the test and criterion measures were obtained (e.g., types of jobs, characteristics of test takers, time interval separating collection of test and criterion measures, year or decade in which the data were collected). If test-criterion relationships vary according to such moderator variables, then, the numbers of studies permitting, the meta-analysis should report separate estimated effect size distributions conditional upon relevant situational features. This might be accomplished, for example, by reporting separate distributions for subsets of studies or by estimating the magnitudes of the influences of situational features on effect sizes.

Standard 1.21

Any meta-analytic evidence used to support an intended test use should be clearly described, including methodological choices in identifying and coding studies, correcting for artifacts, and examining potential moderator variables. Assumptions made in correcting for artifacts such as criterion unreliability and range restriction should be presented, and the consequences of these assumptions made clear.

Comment: Meta-analysis inevitably involves judgments regarding a number of methodological choices. The bases for these judgments should be articulated. In the case of choices involving some degree of uncertainty, such as artifact corrections based on assumed values, the uncertainty should be acknowledged and the degree to which conclusions about validity hinge on these assumptions should be examined and reported.

Standard 1.22

When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

Comment: If it is asserted, for example, that using a given test for employee selection will result in reduced employee errors or training costs, evidence in support of that assertion should be provided. A given claim for the benefits of test use may be supported by logical or theoretical argument as well as empirical data. Due weight should be given to findings in the scientific literature that may be inconsistent with the stated expectation.

Standard 1.23

When a test use or score interpretation is recommended on the grounds that testing or the testing program per se will result in some indirect benefit in addition to the utility of information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Due weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted.

Comment: For example, certain educational testing programs have been advocated on the grounds that they would have a salutary influence on classroom instructional practices or would clarify students' understanding of the kind or level of achievement they were expected to attain. To the extent that such claims enter into the justification for a testing program, they become part of the validity argument for test use and so should be examined as part of the validation effort. Due weight should be given to evidence against such predictions, for example, evidence that under some conditions educational testing may have a negative effect on classroom instruction.

Standard 1.24

When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or to the test's failure fully to represent the intended construct.

Comment: The validity of test score interpretations may be limited by construct-irrelevant components or construct underrepresentation. When unintended consequences appear to stem, at least in part, from the use of one or more tests, it is especially important to check

that these consequences do not arise from such sources of invalidity. Although group differences, in and of themselves, do not call into question the validity of a proposed interpretation, they may increase the salience of plausible rival hypotheses that should be investigated as part of the validation effort.

MEASUREMENT

Background

A test, broadly defined, is a set of tasks designed to elicit or a scale to describe examinee behavior in a specified domain, or a system for collecting samples of an individual's work in a particular area. Coupled with the device is a scoring procedure that enables the examiner to quantify, evaluate, and interpret the behavior or work samples. *Reliability* refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups.

The discussion that follows introduces concepts and procedures that may not be familiar to some readers. It is not expected that the brief definitions and explanations presented here will be sufficient to enable the less sophisticated reader to become adequately conversant with these developments. To achieve a better understanding, such readers may need to consult more comprehensive treatments in the measurement literature.

The usefulness of behavioral measurements presupposes that individuals and groups exhibit some degree of stability in their behavior. However, successive samples of behavior from the same person are rarely identical in all pertinent respects. An individual's performances, products, and responses to sets of test questions vary in their quality or character from one occasion to another, even under strictly controlled conditions. This variation is reflected in the examinee's scores. The causes of this variability are generally unrelated to the purposes of measurement. An examinee may try harder, may make luckier guesses, be more alert, feel less anxious, or enjoy better health on one occasion than another. An examinee may have knowledge, experience, or understanding that is more relevant to some tasks than to others in the domain sampled by the test. Some individuals may exhibit less variation in their scores than others, but no examinee is completely consistent. Because of this variation and, in some instances, because of subjectivity in the scoring process, an individual's obtained score and the average score of a group will always reflect at least a small amount of measurement error.

To say that a score includes a component of error implies that there is a hypothetical error-free value that characterizes an examinee at the time of testing. In classical test theory this error-free value is referred to as the person's *true score* for the test or measurement procedure. It is conceptualized as the hypothetical average score resulting from many repetitions of the test or alternate forms of the instrument. In statistical terms, the true score is a personal parameter and each observed score of an examinee is presumed to estimate this parameter. Under an approach to reliability estimation known as *generalizability theory*, a comparable concept is referred to as an examinee's *universe score*. Under *item response theory (IRT)*, a closely related concept is called an examinee's *ability or trait parameter*, though observed scores and trait parameters may be stated in different units. The hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true or universe score for the procedure is called *measurement error*.

The definition of what constitutes a standardized test or measurement procedure has broadened significantly in recent years. At one time the cardinal features of most standardized tests were consistency of the test materials from examinee to examinee, close adherence to stipulated procedures for test administration, and use of prescribed scoring rules that could be applied with a high degree of consistency. These features were, in fact, what made a test "standardized," and they made meaningful norms possible. In employ-

ment settings and certification programs, flexible measurement procedures have been in use for many years. Individualized oral examinations, simulations, analyses of extended case reports, and performance in real-life settings such as clinics are now commonplace. In education, however, large-scale testing programs with a high degree of flexibility in test format and administrative procedures are a relatively recent development. In some programs cumulative portfolios of student work have been substituted for more traditional end-of-year tests of achievement. Other programs now allow examinees to choose their own topics to demonstrate their abilities. Still others permit or encourage small groups of examinees to work cooperatively in completing the test. A science examination, for example, might involve a team of high school students who conduct a study of the sources of pollution in local streams and prepare a report on their findings. Examinations of this kind raise complex issues regarding the domain represented by the test and about the generalizability of individual and group scores. Each step toward greater flexibility almost inevitably enlarges the scope and magnitude of measurement error. However, it is possible that some of the resultant sacrifices in reliability may reduce construct irrelevance or construct underrepresentation in an assessment program.

Characteristics and Implications of Measurement Error

Errors of measurement are generally viewed as random and unpredictable. They are conceptually distinguished from systematic errors, which may also affect performance of individuals or groups, but in a consistent rather than a random manner. For example, a systematic group error would occur as a result of differences in the difficulty of test forms that have not been adequately equated. When one test form is less difficult than another, examinees who take the easier form may be expected to earn a higher average score than those who take the more difficult form. Such a difference would not be considered an error of measurement under most methods of quantifying and summarizing error, though generalizability theory would permit test form differences to be recognized as an error source.

The systematic factors that may differentially affect the performance of individual test takers are not as easily detected or overridden as those affecting groups. For example, some examinees experience levels of test anxiety that severely impair cognitive efficiency. The presence of such a condition can sometimes be recognized in an examinee, but the effect cannot be overcome by statistical adjustments. The individual systematic errors are not generally regarded as an element that contributes to unreliability. Rather, they constitute a source of construct-irrelevant variance and thus may detract from validity.

Important sources of measurement error may be broadly categorized as those rooted within the examinees and those external to them. Fluctuations in the level of an examinee's motivation, interest, or attention and the inconsistent application of skills are clearly internal factors that may lead to score inconsistencies. Differences among testing sites in their freedom from distractions, the random effects of scorer subjectivity, and variation in scorer standards are examples of external factors. The potency and importance of any particular source depend on the specific conditions under which the measures are taken, how performances are scored, and the interpretations made from the scores. A particular factor, such as the subjectivity in scoring, may be a significant source of measurement error in some assessments and a minor consideration in others.

Some changes in scores from one occasion to another, it should be noted, are not regarded as error, because they result, in part, from an intervention, learning, or maturation that has occurred between the initial and final measures. The difference within an individual indicates, to some extent, the effects of the intervention or the extent of growth. In such settings, change per se constitutes the phenomenon of interest. The difference or the change score then becomes the measure to which reliability pertains.

Measurement error reduces the usefulness of measures. It limits the extent to which test results can be generalized beyond the particulars of a specific application of the measurement process. Therefore, it reduces the confidence that can be placed in any single measurement. Because random measurement errors are inconsistent and unpredictable, they cannot be removed from observed scores. However, their aggregate magnitude can be summarized in several ways, as discussed below.

Summarizing Reliability Data

Information about measurement error is essential to the proper evaluation and use of an instrument. This is true whether the measure is based on the responses to a specific set of questions, a portfolio of work samples, the performance of a task, or the creation of an original product. The ideal approach to the study of reliability entails independent replication of the entire measurement process. However, only a rough or partial approximation of such replication is possible in many testing situations, and investigation of measurement error may require special studies that depart from routine testing procedures. Nevertheless, it should be the goal of test developers to investigate test reliability as fully as practical considerations permit. No test developer is exempt from this responsibility.

The critical information on reliability includes the identification of the major sources of error, summary statistics bearing on the size of such errors, and the degree of generalizability of scores across alternate forms, scorers, administrations, or other relevant dimensions. It also includes a description of the examinee population to whom the foregoing data apply, as the data may accurately reflect what is true of one population but misrepresent what is true of another. For example, a given reliability coefficient or estimated standard error derived from scores of a nationally representative sample may differ significantly from that obtained for a more homogeneous sample drawn from one gender, one ethnic group, or one community.

Reliability information may be reported in terms of variances or standard deviations of measurement errors, in terms of one or more coefficients, or in terms of IRT-based test information functions. The standard error of measurement is the standard deviation of a hypothetical distribution of measurement errors that arises when a given population is assessed via a particular test or procedure. The overall variance of measurement errors is actually a weighted average of the values that hold at various true score levels. The variance at a particular level is called a *conditional error variance* and its square root a *conditional standard error*. Traditionally, three broad categories of reliability coefficients have been recognized: (a) coefficients derived from the administration of parallel forms in independent testing sessions (alternate-form coefficients); (b) coefficients obtained by administration of the same instrument on separate occasions (test-retest or stability coefficients); and (c) coefficients based on the relationships among scores derived from individual items or subsets of the items within a test, all data accruing from a single administration (internal consistency coefficients). Where test scoring involves a high level of judgment, indexes of scorer consistency are commonly obtained. With the development of generalizability theory, the foregoing three categories may now be seen as special cases of a more general classification: generalizability coefficients.

Like traditional reliability coefficients, a *generalizability coefficient* is defined as the ratio of true or universe score variance to observed score variance. Unlike traditional approaches to the study of reliability, however, generalizability theory permits the researcher to specify and estimate the various components of true score variance, error variance, and observed score variance. Estimation is typically accomplished by the application of the techniques of analysis of variance. Of special interest are the separate numerical estimates of the components of overall error variance. Such estimates permit examination of the contribution of each source of error to the overall measurement process. The generalizability approach also makes possible the estimation of coefficients that apply to a wide variety of potential measurement designs.

The test information function, an important result of IRT, efficiently summarizes how well the test discriminates among individuals at various levels of the ability or trait being assessed. Under the IRT conceptualization, a mathematical function called the *item characteristic curve* or *item response function* is used as a model to represent the increasing proportion of correct responses to an item for groups at progressively higher levels of the ability or trait being measured. Given an adequate database, the parameters of the characteristic curve of each item in a test can be estimated. The test information function can then be approximated. This function may be viewed as a mathematical statement of the precision of measurement at each level of the given trait. Precision, in the IRT context, is analogous to the reciprocal of the conditional error variance of classical test theory.

Interpretation of Reliability Data

In general, reliability coefficients are most useful in comparing tests or measurement procedures, particularly those that yield scores in different units or metrics. However, such comparisons are rarely straightforward. Allowance must be made for differences in the variability of the groups on which the coefficients are based, the techniques used to obtain the coefficients, the sources of error reflected in the coefficients, and the lengths of the instruments being compared in terms of testing time.

Generalizability coefficients and the many coefficients included under the traditional categories may appear to be interchangeable, but some convey quite different information from others. A coefficient in any given category may encompass errors of measurement from a highly restricted perspective, a very broad perspective, or some point between these extremes. For example, a coefficient may reflect error due to scorer inconsistencies but not reflect the variation that characterizes a succession of examinee performances or products. A coefficient may reflect only the internal consistency of item responses within an instrument and fail to reflect measurement error associated with day-to-day changes in examinee health, efficiency, or motivation.

It should not be inferred, however, that alternate-form or test-retest coefficients based on test administrations several days or weeks apart are always preferable to internal consistency coefficients. For many tests, internal consistency coefficients do not differ significantly from alternate-form coefficients. Where only one form of a test exists, retesting may result in an inflated correlation between the first and second scores due to idiosyncratic features of the test or to examinee recall of initial responses. Also, an individual's status on some attributes, such as mood or emotional state, may change significantly in a short period of time. In the assessment of such constructs the multiple measures that give rise to reliability estimates should be obtained within the short period in which the attribute remains stable. Therefore, for characteristics of this kind an internal consistency coefficient may be preferred.

The standard error of measurement is generally more relevant than the reliability coefficient once a measurement procedure has been adopted and interpretation of scores has become the user's primary concern. It should be noted that standard errors share some of the ambiguities which characterize reliability coefficients, and estimates may vary in their quality. Information about the precision of measurement at each of several widely spaced score levels—that is, conditional standard errors—is usually a valuable supplement to the single statistic for all score levels combined. Like reliability and generalizability coefficients, standard errors may reflect variation from many sources of error or only a few. For most purposes, a more comprehensive standard error is more informative than a less comprehensive value. However, there are many exceptions to this generalization. Practical constraints often preclude conduct of the kinds of studies that would yield estimates of the preferred standard errors.

Measurements derived from observations of behavior or evaluations of products are especially sensitive to a variety of error factors. These include evaluator biases and idiosyncrasies, scoring subjectivity, and intra-examinee factors that cause variation from one performance or product to another. The methods of generalizability theory are well suited to the investigation of the reliability of the scores on such measures. Estimates of the error variance associated with each specific source and with the interactions between sources indicate the extent to which examinee scores may be generalized to a population of scorers and to a universe of products or performances.

The interpretations of test scores may be broadly categorized as *relative* or *absolute*. Relative interpretations convey the standing of an individual or group within a reference population. Absolute interpretations relate the status of an individual or group to defined standards. These standards may originate in empirical data for one or more populations or be based entirely on authoritative judgment. Different values of the standard error apply to the two types of interpretations.

The test information function can be perceived an alternative to traditional indices of measurement precision, but there are important distinctions that should be noted. Standard errors under classical test theory can be derived by several different approaches. These yield similar, but not identical, results. More significantly, standard errors, like reliability coefficients, may reflect a broad configuration of error factors or a restricted configuration, depending on the design of the reliability study. Test information functions, on the other hand, are limited to the restricted definition of measurement error that is associated with internal consistency reliabilities. In addition, under IRT several different mathematical models have been proposed and accepted as the basic form of the item characteristic curve. Adoption of one model rather than another can have a material effect on the derived test information function.

A final consideration has significant implications for both IRT and classical approaches to quantification of test score precision. It is this: Indices of precision depend on the scale in which they are reported. An index stated in terms of raw scores or the trait level estimates of IRT may convey a radically different perception of reliability than the same index restated in terms of derived scores. This same contrast may hold for conditional standard errors. In terms of the basic score scale, precision may appear to be high at one score level, low at another. But when the conditional standard errors are restated in units of derived scores, such as grade equivalents or standard scores, quite different trends in comparative precision may emerge. Therefore, measurement precision under both theories very strongly depends on the scale in which test scores are reported and interpreted.

Precision and consistency in measurement are always desirable. However, the need

for precision increases as the consequences of decisions and interpretations grow in importance. If a decision can and will be corroborated by information from other sources or if an erroneous initial decision can be quickly corrected, scores with modest reliability may suffice. But if a test score leads to a decision that is not easily reversed, such as rejection or admission of a candidate to a professional school or the decision by a jury that a serious injury was sustained, the need for a high degree of precision is much greater.

Where the purpose of measurement is classification, some measurement errors are more serious than others. An individual who is far above or far below the value established for pass/fail or for eligibility for a special program can be mismeasured without serious consequences. Mismeasurement of examinees whose true scores are close to the cut score is a more serious concern. The techniques used to quantify reliability should recognize these circumstances. This can be done by reporting the conditional standard error in the vicinity of the critical value.

Some authorities have proposed that a semantic distinction be made between "reliability of scores" and "degree of agreement in classification." The former term would be reserved for analysis of score variation under repeated measurement. The term *classification consistency* or *inter-rater agreement*, rather than *reliability*, would be used in discussions of consistency of classification. Adoption of such usage would make it clear that the importance of an error of any given size depends on the proximity of the examinee's score to the cut score. However, it should be recognized that the degree of consistency or agreement in examinee classification is specific to the cut score employed and its location within the score distribution.

Average scores of groups, when interpreted as measures of program effectiveness, involve error factors that are not identical to those that operate at the individual level. For large groups, the positive and negative measurement errors of individuals may average out almost completely in group means. However, the sampling errors associated with the random sampling of persons who are tested for purposes of program evaluation are still present. This component of the variation in the mean achievement of school classes from year to year or in the average expressed satisfaction of successive samples of the clients of a program may constitute a potent source of error in program evaluations. It can be a significant source of error in inferences about programs even if there is a high degree of precision in individual test scores. Therefore, when an instrument is used to make group judgments, reliability data must bear directly on the interpretations specific to groups. Standard errors appropriate to individual scores are not appropriate measures of the precision of group averages. A more appropriate statistic is the standard error of the observed score means. Generalizability theory can provide more refined indices when the sources of measurement error are numerous and complex.

Typically, developers and distributors of tests have primary responsibility for obtaining and reporting evidence of reliability or test information functions. The user must have such data to make an informed choice among alternative measurement approaches and will generally be unable to conduct reliability studies prior to operational use of an instrument. In some instances, however, local users of a test or procedure must accept at least partial responsibility for documenting the precision of measurement. This obligation holds when one of the primary purposes of measurement is to rank or classify examinees within the local population. It also holds when users must rely on local scorers who are trained to use the scoring rubrics provided by the test developer. In such settings, local factors may materially affect the magnitude of error variance and observed score variance. Therefore, the reliability of scores may differ appreciably from that reported by the developer.

The reporting of reliability coefficients alone, with little detail regarding the methods used to estimate the coefficient, the nature of the group from which the data were derived, and the conditions under which the data were obtained constitutes inadequate documentation. General statements to the effect that a test is "reliable" or that it is "sufficiently reliable to permit interpretations of individual scores" are rarely, if ever, acceptable. It is the user who must take responsibility for determining whether or not scores are sufficiently trustworthy to justify anticipated uses and interpretations. Of course, test constructors and publishers are obligated to provide sufficient data to make informed judgments possible.

As the foregoing comments emphasize, there is no single, preferred approach to quantification of reliability. No single index adequately conveys all of the relevant facts. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment.

Although reliability is discussed here as an independent characteristic of test scores, it should be recognized that the level of reliability of scores has implications for the validity of score interpretations. Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores. The data also bear on the consistency of classifications of individuals derived from the scores. To the extent that scores reflect random errors of measurement, their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited. Relatively unreliable scores, in conjunction with other convergent information, may sometimes be of value to a test user, but the level of a score's reliability places limits on its unique contribution to validity for all purposes.

Standard 2.1

For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.

Comment: It is not sufficient to report estimates of reliabilities and standard errors of measurement only for total scores when subscores are also interpreted. The form-to-form and day-to-day consistency of total scores on a test may be acceptably high, yet subscores may have unacceptably low reliability. For all scores to be interpreted, users should be supplied with reliability data in enough detail to judge whether scores are precise enough for the users' intended interpretations. Composites formed from selected subtests within a test battery are frequently proposed for predictive and diagnostic purposes. Users need information about the reliability of such composites.

Standard 2.2

The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.

Comment: The most common derived scores include standard scores, grade or age equivalents, and percentile ranks. Because raw scores on norm-referenced tests are only rarely interpreted directly, standard errors in derived score units are more helpful to the typical test user. A confidence interval for an examinee's true score, universe score, or percentile rank serves much the same purpose as a standard error and can be used as an alternative approach to convey reliability information. The implications of the standard error of measurement are especially important in situations where decisions cannot be postponed and corroborative sources of information are limited.

Standard 2.3

When test interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability data, including standard errors, should be provided for such differences.

Comment: Observed score differences are used for a variety of purposes. Achievement gains are frequently the subject of inferences for groups as well as individuals. Differences between verbal and performance scores of intelligence and scholastic ability tests are often employed in the diagnosis of cognitive impairment and learning problems. Psychodiagnostic inferences are frequently drawn from the differences between subtest scores. Aptitude and achievement batteries, interest inventories, and personality assessments are commonly used to identify and quantify the relative strengths and weaknesses or the pattern of trait levels of an examinee. When the interpretation of test scores centers on the peaks and valleys in the examinee's test score profile, the reliability of score differences for all pairs of scores is critical.

Standard 2.4

Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported.

Comment: Information on the method of subject selection, sample sizes, means, standard deviations, and demographic characteristics of the groups helps users judge the extent to which reported data apply to their own examinee populations. If the test-retest or alternate-form approach is used, the interval between testings should be indicated. Because there are many ways of estimating reliability, each influenced by different sources of measurement error, it is unacceptable to say simply, "The reliability of test X is .90." A better statement would be, "The reliability coefficient of .90 reported for scores on test X was obtained by correlating scores from forms A and B administered on successive days. The data were based on a sample of 400 10th-grade students from five middle-class suburban schools in New York State. The demographic breakdown of this group was as follows: ...."

Standard 2.5

A reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent.

Comment: Internal consistency, alternate-form, test-retest, and generalizability coefficients should not be considered equivalent, as each may incorporate a unique definition of measurement error. Error variances derived via item response theory may not be equivalent to error variances estimated via other approaches. Test developers should indicate the sources of error that are reflected in or ignored by the reported reliability indices.

Standard 2.6

If reliability coefficients are adjusted for restriction of range or variability, the adjustment procedure and both the adjusted and unadjusted coefficients should be reported. The standard deviations of the group actually tested and of the target population, as well as the rationale for the adjustment, should be presented.

Comment: Application of a correction for restriction in variability presumes that the available sample is not representative of the test-taker population to which users might be expected to generalize. The rationale for the correction should consider the appropriateness of such a generalization. Adjustment formulas that presume constancy in the standard error across score levels should not be used unless constancy can be defended.

Standard 2.7

When subsets of items within a test are dictated by the test specifications and can be presumed to measure partially independent traits or abilities, reliability estimation procedures should recognize the multifactor character of the instrument.

Comment: The total score on a test that is clearly multifactor in nature should be treated as a composite score. If an internal consistency estimate of total score reliability is obtained by the split-halves procedure, the halves should be parallel in content and statistical characteristics. Stratified coefficient alpha should be used rather than the more familiar unstratified coefficient.

Standard 2.8

Test users should be informed about the degree to which rate of work may affect examinee performance.

Comment: It is not possible to state, in general, whether reliability coefficients will increase or decrease when rate of work becomes an important source of systematic variance. Rate of work, as an examinee trait, may be more stable or less stable from occasion to occasion than the other factors the test is designed to measure. Because speededness has differential effects on various estimates, information on speededness is helpful in interpreting reported coefficients.

The importance of the speed factor can sometimes be inferred from analyses of item responses and from observations by examiners during test administrations conducted for reliability analyses. The distribution of "last item attempted" and increases in the frequency of omitted responses toward the end of a test are also highly informative, though not conclusive, evidence regarding speededness. A decline in the proportion of correct responses, beyond that attributable to increasing item difficulty, may indicate that some examinees were responding randomly. With computer-administered tests, abnormally fast item response times, particularly toward the end of the test, may also suggest that examinees were responding randomly. In the case of constructed-response exercises, including essay questions, the completeness of the responses may suggest that time constraints had little effect on early items but a significant effect on later items. Introduction of a speed factor into what might otherwise be a power test may have a marked effect on alternate-form and test-retest reliabilities. A shift from a paper-and-pencil format to a computer-administered format may affect test speededness.

Standard 2.9

When a test is designed to reflect rate of work, reliability should be estimated by the alternate-form or test-retest approach, using separately timed administrations.

Comment: Split-half coefficients based on separate scores from the odd-numbered and even-numbered items are known to yield inflated estimates of reliability for highly speeded tests. Coefficient alpha and other internal consistency coefficients may also be biased, though the size of the bias is not as clear as that for the split-halves coefficient.

Standard 2.10

When subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same perform-

successive performances or new products, and (c) independent panels scoring successive performances or new products.

Comment: Task-to-task variations in the quality of an examinee's performance and rater-to-rater inconsistencies in scoring represent independent sources of measurement error. Reports of reliability studies should make clear which of these sources are reflected in the data. Where feasible, the error variances arising from each source should be estimated. Generalizability studies and variance component analyses are especially helpful in this regard. These analyses can provide separate error variance estimates for tasks within examinees, for judges, and for occasions within the time period of trait stability. Information should be provided on the qualifications of the judges used in reliability studies.

Inter-rater or inter-observer agreement may be particularly important for ratings and observational data that involve subtle discriminations. It should be noted, however, that when raters evaluate positively correlated characteristics, a favorable or unfavorable assessment of one trait may color their opinions of other traits. Moreover, high inter-rater consistency does not imply high examinee consistency from task to task. Therefore, internal consistency within raters and inter-rater agreement do not guarantee high reliability of examinee scores.

Standard 2.11

If there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard errors of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended.

Comment: If test score interpretation involves inferences within subpopulations as well as within the general population, reliability data should be provided for both the subpopulations and the general population. Test users who work exclusively with a specific cultural group or with individuals who have a particular disability would benefit from an estimate of the standard error for such a subpopulation. Some groups of test takers—pre-school children, for example—tend to respond to test stimuli in a less consistent fashion than do older children.

Standard 2.12

If a test is proposed for use in several grades or over a range of chronological age groups and if separate norms are provided for each grade or each age group, reliability data should be provided for each age or grade population, not solely for all grades or ages combined.

Comment: A reliability coefficient based on a sample of examinees spanning several grades or a broad range of ages in which average scores are steadily increasing will generally give a spuriously inflated impression of reliability. When a test is intended to discriminate within age or grade populations, reliability coefficients and standard errors should be reported separately for each population.

Standard 2.13

If local scorers are employed to apply general scoring rules and principles specified by the test developer, local reliability data should be gathered and reported by local authorities when adequate size samples are available.

Comment: For example, many statewide testing programs depend on local scoring of essays, constructed-response exercises, and performance tests. Reliability analyses bear on the possibility that additional training of scorers is needed and, hence, should be an integral part of program monitoring.

Standard 2.14

Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Comment: Estimation of conditional standard errors is usually feasible even with the sample sizes that are typically used for reliability analyses. If it is assumed that the standard error is constant over a broad range of score levels, the rationale for this assumption should be presented.

Standard 2.15

When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument.

Comment: When a test or composite is used to make categorical decisions, such as pass/fail, the standard error of measurement at or near the cut score has important implications for the trustworthiness of these decisions. However, the standard error cannot be translated into the expected percentage of consistent decisions unless assumptions are made about the form of the distributions of measurement errors and true scores. It is preferable that this percentage be estimated directly through the use of a repeated-measurements approach if consistent with the requirements of test security and if adequate samples are available.

Standard 2.16

In some testing situations, the items vary from examinee to examinee—through random selection from an extensive item pool or application of algorithms based on the examinee's level of performance on previous items or preferences with respect to item difficulty. In this type of testing, the preferred approach to reliability estimation is one based on successive administrations of the test under conditions similar to those prevailing in operational test use.

Comment: Varying the set of items presented to each examinee is an acceptable procedure in some settings. If this approach is used, reliability data should be appropriate to this procedure. Estimates of standard errors of ability scores can be computed through the use of IRT and reported routinely as part of the adaptive testing procedure. However, those estimates are not an adequate substitute for estimates based on successive administrations of the adaptive test, nor do they bear on the issue of stability over short intervals. IRT estimates are contingent on the adequacy of both the item parameter estimates and the item response models adopted in the theory. Estimates of reliabilities and standard errors of measurement based on the administration and analysis of alternate forms of an adaptive test reflect errors associated with the entire measurement process. The alternate-form estimates provide an independent check on the magnitude of the errors of measurement specific to the adaptive feature of the testing procedure.

Standard 2.17

When a test is available in both long and short versions, reliability data should be reported for scores on each version, preferably based on an independent administration of each.

Comment: Some tests and test batteries are published in both a "full-length" version and a "survey" or "short" version. In many applications the Spearman-Brown formula will satisfactorily approximate the reliability of one of these from data based on the other. However, context effects are commonplace in tests of

maximum performance. Also, the short version of a standardized test often comprises a nonrandom sample of items from the full-length version. Therefore, the shorter version may be more reliable or less reliable than the Spearman-Brown projections from the full-length version. The reliability of scores on each version is best evaluated through an independent administration of each, using the designated time limits.

Standard 2.18

When significant variations are permitted in test administration procedures, separate reliability analyses should be provided for scores produced under each major variation if adequate sample sizes are available.

Comment: To accommodate examinees with disabilities, test publishers might authorize modifications in the procedures and time limits that are specified for the administration of the paper-and-pencil edition of a test. In some cases, modified editions of the test itself may be provided. For example, tape-recorded versions for use in a group setting or with individual equipment may be used to test examinees who exhibit reading disabilities or attention deficits. If such modifications can be employed with test takers who are not disabled, insights can be gained regarding the possible effects on test scores of these nonstandard administrations.

Standard 2.19

When average test scores for groups are used in program evaluations, the groups tested should generally be regarded as a sample from a larger population, even if all examinees available at the time of measurement are tested. In such cases the standard error of the group mean should be reported, as it reflects variability due to sampling of examinees as well as variability due to measurement error.

Comment: The graduating seniors of a liberal arts college, the current clients of a social service agency, and analogous groups exposed to a program of interest typically constitute a sample in a longitudinal sense. Presumably, comparable groups from the same population will recur in future years, given static conditions. The factors leading to uncertainty in conclusions about program effectiveness arise from the sampling of persons as well as measurement error. Therefore, the standard error of the mean observed score, reflecting variation in both true scores and measurement errors, represents a more realistic standard error in this setting. Even this value may underestimate the variability of group means over time. In many settings, the static conditions assumed under random sampling of persons do not prevail.

Standard 2.20

When the purpose of testing is to measure the performance of groups rather than individuals, a procedure frequently used is to assign a small subset of items to each of many subsamples of examinees. Data are aggregated across subsamples and item subsets to obtain a measure of group performance. When such procedures are used for program evaluation or population descriptions, reliability analyses must take the sampling scheme into account.

Comment: This type of measurement program is termed *matrix sampling*. It is designed to reduce the time demanded of individual examinees and to increase the total number of items on which data are obtained. This testing approach provides the same type of information about group performances that would accrue if all examinees could respond to all exercises in the item pool. Reliability statistics must be appropriate to the sampling plan used with respect to examinees and items.

Background

Test development is the process of producing a measure of some aspect of an individual's knowledge, skill, ability, interests, attitudes, or other characteristics by developing items and combining them to form a test, according to a specified plan. Test development is guided by the stated purpose(s) of the test and the intended inferences to be made from the test scores. The test development process involves consideration of content, format, the context in which the test will be used, and the potential consequences of using the test. Test development also includes specifying conditions for administering the test, determining procedures for scoring the test performance, and reporting the scores to test takers and test users. This chapter focuses primarily on the following aspects of test development: stating the purpose(s) of the test, defining a framework for the test, developing test specifications, developing and evaluating items and their associated scoring procedures, assembling the test, and revising the test. The first section describes the test development process that begins with a statement of the purpose(s) of the test and culminates with the assembly of the test. The second section addresses several special considerations in test development, including considerations in delineating the test framework and in developing performance assessments. The chapter concludes with a discussion on test revision. Issues bearing on validity, reliability, and fairness are interwoven within the stages of test development. Each of these topics is addressed comprehensively in other chapters of the *Standards*: validity in chapter 1, reliability in chapter 2, and aspects of fairness in chapters 7, 8, 9, and 10. Additional material on test administration and scoring, and on reporting scores and results, is provided in chapter 5. Chapter 4 discusses score scales, and the focus of chapter 6 is test documents.

Test Development

The process of developing educational and psychological tests commonly begins with a statement of the purpose(s) of the test and the construct or content domain to be measured. Tests of the same construct or domain can differ in important ways, because a number of decisions must be made as the test is developed. It is helpful to consider the four phases leading from the original statement of purpose(s) to the final product: (a) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; (b) development and evaluation of the test specifications; (c) development, field testing, evaluation, and selection of the items and scoring guides and procedures; and (d) assembly and evaluation of the test for operational use. What follows is a description of typical test development procedures, though there may be sound reasons that some of these steps are followed in some settings and not in others.

The first step is to extend the original statement of purpose(s), and the construct or content domain being considered, into a framework for the test that describes the extent of the domain, or the scope of the construct to be measured. The test framework, therefore, delineates the aspects (e.g., content, skills, processes, and diagnostic features) of the construct or domain to be measured. For example, "Does eighth-grade mathematics include algebra?" "Does verbal ability include text comprehension as well as vocabulary?" "Does self-esteem include both feelings and acts?" The delineation of the test framework can be guided by theory or an analysis of the content domain or job requirements as in the case of many licensing and employment tests. The test framework serves as a guide to subsequent test evaluation. The chapter on validity provides a more thorough discussion of the relationships among the construct or content domain, the test framework, and the purpose(s) of the test.

Once decisions have been made about the test is to measure, and what its scores ntended to convey, the next step is to n the test by establishing test specifica- The test specifications delineate the for- f items, tasks, or questions; the response at or conditions for responding; and the of scoring procedures. The specifications indicate the desired psychometric prop- of items, such as difficulty and discrimi- n, as well as the desired test properties as test difficulty, inter-item correlations, reliability. The test specifications may include such factors as time restrictions, acteristics of the intended population of akers, and procedures for administration. ubsequent test development activities are ed by the test specifications.

Test specifications will include, at least licitly, an indication of whether the test es will be primarily norm-referenced or rion-referenced. When scores are norm- renced, relative score interpretations are of nary interest. A score for an individual or a definable group is ranked within one or re distributions of scores or compared to average performance of test takers for var- reference populations (e.g., based on age, de, diagnostic category, or job classifica-). When scores are criterion-referenced, olute score interpretations are of primary rest. The meaning of such scores does not end on rank information. Rather, the test re conveys directly a level of competence ome defined criterion domain. Both rela- e and absolute interpretations are often d with a given test, but the test developer ermines which approach is most relevant that test.

The nature of the item and response for- ts that may be specified depends on the rposes of the test and the defined domain the test. Selected-response formats, such as ultiple-choice items, are suitable for many rposes of testing. The test specifications dicate how many alternatives are to be used for each item. Other purposes may be more effectively served by a short constructed-response format. Short-answer items require a response of no more than a few words. Extended-response formats require the test taker to write a more extensive response of one or more sentences or paragraphs. Performance assessments often seek to emulate the context or conditions in which the intended knowledge or skills are actually applied. One type of performance assessment, for example, is the standardized job or work sample. A task is presented to the test taker in a standardized format under standardized conditions. Job or work samples might include, for example, the assessment of a practitioner's ability to make an accurate diagnosis and recommend treatment for a defined condition, a manager's ability to articulate goals for an organization, or a student's proficiency in performing a science laboratory experiment.

All types of items require some indication of how to score the responses. For selected-response items, one alternative is considered the correct response in some testing programs. In other testing programs, the alternatives may be weighted differentially. For short-answer items, a list of acceptable alternatives may suffice; extended-response items need more detailed rules for scoring, sometimes called *scoring rubrics*. Scoring rubrics specify the criteria for evaluating performance and may vary in the degree of judgment entailed, in the number of score levels, and in other ways. It is common practice for test developers to provide scorers with examples of performances at each of the score levels to help clarify the criteria.

For extended-response items, including performance tasks, two major types of scoring procedures are used: analytic and holistic. Both of the procedures require explicit performance criteria that reflect the test framework. However, the approaches differ in the degree of detail provided in the evaluation report. Under the analytic scoring procedure, each critical dimension of the performance criteria is judged independently, and separate scores are obtained for each of these dimensions in addition to an overall score. Under the holistic scoring procedure, the same performance criteria may implicitly be considered, but only one overall score is provided. Because the analytic procedure provides information on a number of critical dimensions, it potentially provides valuable information for diagnostic purposes and lends itself to evaluating strengths and weaknesses of test takers. In contrast, the holistic procedure may be preferable when an overall judgment is desired and when the skills being assessed are complex and highly interrelated. Regardless of the type of scoring procedure, designing the items and developing the scoring rubrics and procedures is an integrated process.

A participatory approach may be used in the design of items, scoring rubrics, and sometimes the scoring process itself. Many interested persons (e.g., practitioners, teachers) may be involved in developing items and scoring rubrics, and/or evaluating the subsequent performances. If a participatory approach is used, participants' knowledge about the domain being assessed and their ability to apply the scoring rubrics are of critical importance. Equally important, for those involved in developing tests and evaluating performances, is their familiarity with the nature of the population being tested. Relevant characteristics of the population being tested may include the typical range of expected skill levels, their familiarity with the response modes required of them, and the primary language they use.

The test developer usually assembles an item pool that consists of a larger set of items than what is required by the test specifications. This allows for the test developer to select a set of items for the test that meet the test specifications. The quality of the items is usually ascertained through item review procedures and pilot testing. Items are reviewed for content quality, clarity and lack of ambiguity. Items sometimes are reviewed for sensitivity to gender or cultural issues. An attempt is generally made to avoid words and topics that may offend or otherwise disturb some test takers, if less offensive material is equally useful. Often, a field test is developed and administered to a group of test takers who are somewhat representative of the target population for the test. The field test helps determine some of the psychometric properties of the test items, such as an item's difficulty and ability to discriminate among test takers of different standing on the scale. Ongoing testing programs often pretest items by inserting them into existing tests. Those items are not used in obtaining test scores of the test takers, but the item responses provide useful data for test development.

The next step in test development is to assemble items into a test or so identify an item pool for an adaptive test. The test developer is responsible for ensuring that the items selected for the test meet the requirements of the test specifications. Depending upon the purpose(s) of the test, relevant considerations in item selection may include the content quality and scope, the weighting of items and subdomains, and the appropriateness of the items selected for the intended population of test takers. Often test developers will specify the distribution of psychometric indices of the items to be included in the test. For example, the specified distribution of item difficulty indices for a selection test would differ from the distribution specified for a general achievement test. When psychometric indices of the items are estimated using item response theory (IRT), the fit of the model to the data is also evaluated. This is accomplished by evaluating the extent to which the assumptions underlying the item response model (e.g., unidimensionality, local independence, speededness, and equality of slope parameters) are satisfied.

The test developer is also responsible for ensuring that the scoring procedures are consistent with the purpose(s) of the test and facilitate meaningful score interpretation. The nature of the intended score interpretations

determine the importance of psychometric acteristics of items in the test construction ess. For example, indices of item difficulty discrimination, and inter-item correlations, be particularly important when relative e interpretations are intended. In the case elative score interpretations, good discrim- ion among test takers at all points along construct continuum is desirable. It is portant, however, that the test specifica- is are not compromised when optimizing distribution of these indices. In the case bsolute score interpretations, different cri- a apply. In this case, the extent to which relevant domain has been adequately rep- ented is important even if many of the ms are relatively easy or nondiscriminating thin a relevant population. It is important, wever, to assure the quality of the content relatively easy or nondiscriminating items. cut scores are necessary for score interpreta- n in criterion-referenced programs, the level item discrimination constitutes critical formation primarily in the vicinity of the t scores. Because of these differences in test velopment procedures, tests designed to cilitate one type of interpretation function ss effectively for other types of interpretation. iven appropriate test design and supporting vidence, however, scores arising from some orm-referenced programs may provide rea- onable absolute score interpretations and cores arising from some criterion-refer- nced programs may provide reasonable rela- ive score interpretations.

When evaluating the quality of the items n the item pool and the test itself, test devel- pers often conduct studies of differential tem functioning (see chapter 7). Differential tem functioning is said to exist when test takers of approximately equal ability on the targeted construct or content domain differ in their responses to an item according to their group membership. In theory, the ultimate goal of such studies is to identify construct- irrelevant aspects of item content, item format, or scoring criteria that may differentially affect test scores of one or more groups of test takers. When differential item functioning is detected, test developers try to identify plausible explanations for the differences, and then they may replace or revise items that give rise to group differences if construct irrelevance is deemed likely. However, at this time, there has been little progress in discerning the cause or substantive themes that account for differential item functioning on a group basis. Items for which the differential item functioning index is significant may constitute valid measures of an element of the intended domain and differ in no way from other items that show nonsignificant indexes. When the differential item functioning index is significant, the test developer must take care that any replacement items or item revisions do not compromise the test specifications.

When multiple forms of a test are prepared, the test specifications govern each of the forms. Also, when an item pool is developed for a computerized adaptive test, the specifications refer both to the item pool and to the rules or procedures by which the individual item sets are created for each test taker. Some of the attractive features of computerized adaptive tests, such as tailoring the difficulty level of the items to the test taker's ability, place additional constraints on the design of such tests. In general, a large number of items is needed for a computerized adaptive test to ensure that each tailored item set meets the requirements of the test specifications. Further, tests often are developed in the context of larger systems or programs. Multiple item sets, for example, may be created for use with different groups of test takers or on different testing dates. Last, when a short form of a test is prepared, the test specifications of the original test govern the short form. Differences in the test specifications and the psychometric properties of the short form and the original test will affect the interpretation of the scores derived from the short

form. In any of these cases, the same fundamental methods and principles of test development apply.

Special Considerations in Test Development

This section elaborates on several topics discussed above. First, considerations in delineating the framework for the test are discussed. Following this, considerations in the development of performance assessments and portfolios are addressed.

Delineating the Framework for the Test

The scenario presented above outlines what is often done to develop a test. However, the activities do not always happen in a rigid sequence. There is often a subtle interplay between the process of conceptualizing a construct or content domain and the development of a test of that construct or domain. The framework for the test provides a description of how the construct or domain will be represented. The procedures used to develop items and scoring rubrics and to examine item characteristics may often contribute to clarifying the framework. The extent to which the framework is defined a priori is dependent on the testing application. In many testing applications, a well-defined framework and detailed test specifications guide the development of items and their associated scoring rubrics and procedures. In some areas of psychological measurement, test development may be less dependent on an a priori defined framework and may rely more on a data-based approach that results in an empirically derived definition of the framework. In such instances, items are selected primarily on the basis of their empirical relationship with an external criterion, their relationships with one another, or their power to discriminate among groups of individuals. For example, construction of a selection test for sales personnel might be guided by the correlations of item scores with productivity measures of current sales personnel or a measure of client satisfaction might be assembled from those items in an item pool that correlate most highly with customer loyalty. Similarly, an inventory to help identify different patterns of psychopathology might be developed using patients from different diagnostic subgroups. When test development relies on a data-based approach, it is likely that some items will be selected based on chance occurrences in the data. Cross-validation studies are routinely conducted to determine the tendency to select items by chance, which involves administering the test to a comparable sample.

In many testing applications, the framework for the test is specified initially and this specification subsequently guides the development of items and scoring procedures. Empirical relationships may then be used to inform decisions about retaining, rejecting, or modifying items. Interpretations of scores from tests developed by this process have the advantage of a logical/theoretical and an empirical foundation for the underlying dimensions represented by the test.

Performance Assessments

One distinction between performance assessments and other forms of tests has to do with the type of response that is required from the test takers. Performance assessments require the test takers to carry out a process such as playing a musical instrument or tuning a car's engine or to produce a product such as a written essay. Performance assessments generally require the test takers to demonstrate their abilities or skills in settings that closely resemble real-life settings. For example, an assessment of a psychologist in training may require the test taker to interview a client, choose appropriate tests, and arrive at diagnosis and plan for therapy. Performance assessments are diverse in nature and can be product-based as well as behavior-based. Because performance assessments typically consist of a small num-

ber of tasks, establishing the extent to which the results can be generalized to the broader domain is particularly important. The use of test specifications will contribute to tasks being developed so as to systematically represent the critical dimensions to be assessed, leading to a more comprehensive coverage of the domain than what would occur if test specifications were not used. Further, both logical and empirical evidence are important to document the extent to which performance assessments—tasks as well as scoring criteria—reflect the processes or skills that are specified by the domain definition. When tasks are designed to elicit complex cognitive processes, logical analyses of the tasks and both logical and empirical analyses of the test takers' performances on the tasks provide necessary validity evidence.

Portfolios

A unique type of performance assessment is an individual portfolio. Portfolios are systematic collections of work or educational products typically collected over time. Like other assessment procedures, the design of portfolios is dependent on the purpose. Typical purposes include judgment of the improvement in job or educational performance and evaluation of the eligibility for employment, promotion, or graduation. A well-designed portfolio specifies the nature of the work that is to be put into the portfolio. The portfolio may include entries such as representative products, the best work of the test taker, or indicators of progress. For example, in an employment setting involving promotion, employees may be instructed to include their best work or products. Alternatively, if the purpose is to judge a student's educational growth, students may be asked to provide evidence of improvement with respect to particular competencies or skills. They may also be requested to provide justifications for the choices. Still other methods may include the use of videotapes, exhibitions, demonstrations, simulations, and so on.

In employment settings, employees may be involved in the selection of their work and products that demonstrate their competencies for promotion purposes. Analogously, in educational applications, students may participate in the selection of some of their work and the products to be included in their portfolios as well as in the evaluation of the materials. The specifications for the portfolio indicate who is responsible for selecting its contents. For example, the specifications may state that the test taker, the examiner, or both parties working together should be involved in the selection of the contents of the portfolio. The particular responsibilities of each party are delineated in the specifications. The more standardized the contents and procedures of administration, the easier it is to establish comparability of portfolio-based scores. Regardless of the methods used, all performance assessments are evaluated by the same standards of technical quality as other forms of tests.

Test Revisions

Tests and their supporting documents (e.g., test manuals, technical manuals, user's guides) are reviewed periodically to determine whether revisions are needed. Revisions or amendments are necessary when new research data, significant changes in the domain, or new conditions of test use and interpretation would either improve the validity of interpretations of the test scores or suggest that the test is no longer fully appropriate for its intended use. As an example, tests are revised if the test content or language has become outdated and, therefore, may subsequently affect the validity of the test score interpretations. Revisions to test content are also made to ensure the confidentiality of the test. It should be noted, however, that outdated norms may not have the same implications for revisions as an outdated test. For example, it may be necessary to update the norms for an achievement test after a period of rising or falling achievement in the norming population, or when there are changes in the test-taking population, but the test content itself may continue to be as relevant as it was when the test was developed.

Standard 3.1

Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development.

Standard 3.2

The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent.

Comment: The adequacy and usefulness of test interpretations depend on the rigor with which the purposes of the test and the domain represented by the test have been defined and explicated. The domain definition should be sufficiently detailed and delimited to show clearly what dimensions of knowledge, skill, processes, attitude, values, emotions, or behavior are included and what dimensions are excluded. A clear description will enhance accurate judgments by reviewers and others about the congruence of the defined domain and the test items.

Standard 3.3

The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information.

Comment: Professional judgment plays a major role in developing the test specifications. The specific procedures used for developing the specifications depend on the purposes of the test. For example, in developing licensure and certification tests, practice analyses or job analyses usually provide the basis for defining the test specifications, and job analyses primarily serve this function for employment tests. For achievement tests to be given at the end of a course, the test specifications should be based on an outline of course content and goals. Whereas, for placement tests, it may be necessary to examine the required entry knowledge and skills for several courses.

Standard 3.4

The procedures used to interpret test scores, and, when appropriate, the normative or standardization samples or the criterion used should be documented.

Comment: Test specifications may indicate that the intended score interpretations are for absolute or relative score interpretations, or both. In relative score interpretations the status of an individual (or group) is determined by comparing the score (or mean score) to the performance of others in one or more defined populations. In absolute score interpretations, the score or average is assumed to reflect directly a level of competence or mastery in some defined criterion domain. Tests designed to facilitate one type of interpretation function less effectively for other types of interpretations. Given appropriate test design and adequate supporting data, however, scores arising from norm-referenced testing programs may provide reasonable absolute score interpretations and scores arising from criterion-referenced programs may provide reasonable relative score interpretations.

Standard 3.5

When appropriate, relevant experts external to the testing program should review the test specifications. The purpose of the review, the

process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Comment: Expert review of the test specifications may serve many useful purposes such as helping to assure content quality and representativeness. The expert judges may include individuals representing defined populations of concern to the test specifications. For example, if the test is related to ethnic minority concerns, the expert review typically includes members of appropriate ethnic minority groups or experts on minority group issues.

Standard 3.6

The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Comment: Expert judges may be asked to identify material likely to be inappropriate, confusing, or offensive for groups in the test-taking population. For example, judges may be asked to identify whether lack of exposure to problem contexts in mathematics word problems may be of concern for some groups of students. Various groups of test takers can be defined by characteristics such as age, ethnicity, culture, gender, disability, or demographic region. There is limited evidence, however, that expert reviews alleviate problems with bias in testing (see chapter 7).

Standard 3.7

The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented.

Comment: Empirical evidence and/or expert judgment are used to classify items according to categories of the test specifications. For example, professional panels may be used for classifying the items or for determining the appropriateness of the developer's classification scheme. The panel and procedures used should be chosen with care as they will affect the accuracy of the classification.

Standard 3.8

When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended.

Comment: Conditions which may differentially affect performance on the test items by the sample(s) as compared to the intended population(s) should be documented when appropriate. As an example, test takers may be less motivated when they know their scores will not have an impact on them.

Standard 3.9

When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

Comment: Although overall sample size is important, it is important also that there be an adequate number of cases in regions critical to the determination of the psychometric properties of items. If the test is to achieve greatest precision in a particular part of the score scale and this consideration affects item selection, the manner in which item statistics are used needs to be carefully described. When IRT is used as the basis of test development, it is important to document the adequacy of fit of the model to the data. This is accomplished by providing information about the extent to which IRT assumptions (e.g., unidimensionality, local item independence, or equality of slope parameters) are satisfied.

Test developers should show that any differences between the administration conditions of the field test and the final form do not affect item performance. Conditions that can affect item statistics include item position, time limits, length of test, mode of testing (e.g., paper-and-pencil versus computer-administered), and use of calculators or other tools. For example, in field testing items, those placed at the end of a test might obtain poorer item statistics than those inserted within the test.

Standard 3.10

Test developers should conduct cross-validation studies when items are selected primarily on the basis of empirical relationships rather than on the basis of content or theoretical considerations. The extent to which the different studies identify the same item set should be documented.

Comment: When data-based approaches to test development are used, items are selected primarily on the basis of their empirical relationships with an external criterion, their relationships with one another, or their power to discriminate among groups of individuals. Under these circumstances, it is likely that some items will be selected based on chance occurrences in the data used. Administering the test to a comparable sample of test takers or a hold-out sample provides a means by which the tendency to select items by chance can be determined.

Standard 3.11

Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.

Comment: Test developers should provide evidence of the extent to which the test items and scoring criteria represent the defined domain. This affords a basis to help determine whether performance on the test can be generalized to the domain that is being assessed. This is especially important for tests that contain a small number of items such as performance assessments. Such evidence may be provided by expert judges.

Standard 3.12

The rationale and supporting evidence for computerized adaptive tests should be documented. This documentation should include procedures used in selecting subsets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and for controlling item exposure.

Comment: It is important to assure that documentation of the procedures does not compromise the security of the test items.

If a computerized adaptive test is intended to measure a number of different content subcategories, item selection procedures are to assure that the subcategories are adequately represented by the items presented to the test taker.

Standard 3.13

When a test score is derived from the differential weighting of items, the test developer should document the rationale and process used to develop, review, and assign item weights. When the item weights are obtained based on empirical data, the sample used for obtaining item weights should be sufficiently large and representative of the population for which the test is intended. When the item weights are obtained based on expert judgment, the qualifications of the judges should be documented.

Comment: Changes in the population of test takers, along with other changes such as changes in instructions, training, or job requirements, may impact the original derived item weights, necessitating subsequent studies after an appropriate period of time.

Standard 3.14

The criteria used for scoring test takers' performance on extended-response items should be documented. This documentation is especially important for performance assessments, such as scorable portfolios and essays, where the criteria for scoring may not be obvious to the user.

Comment: The completeness and clarity of the test specifications, including the definition of the domain, are essential in developing the scoring criteria. The test developer needs to provide a clear description of how the test scores are intended to be interpreted to help ensure the appropriateness of the scoring procedures.

Standard 3.15

When using a standardized testing format to collect structured behavior samples, the domain, test design, test specifications, and materials should be documented as for any other test. Such documentation should include a clear definition of the behavior expected of the test takers, the nature of the expected responses, and any materials or directions that are necessary to carry out the testing.

Comment: In developing a prompt, the age, language, experience, and ability level of test takers should be considered, as should other possible unique sources of difficulty for groups in the population to be tested. Test directions that specify time allowances, nature of the responses expected, and rules regarding use of supplementary materials, such as notes, references, dictionaries, calculators, or manipulatives such as lab equipment, may be established via field testing.

Standard 3.16

If a short form of a test is prepared, for example, by reducing the number of items on the original test or organizing portions of a test into a separate form, the specifications of the short form should be as similar as possible to those of the original test. The procedures used for the reduction of items should be documented.

Comment: The extent to which the specifications of the short form differ from those of the original test, and the implications of such differences for interpreting the scores derived from the short form, should be documented.

Standard 3.17

When previous research indicates that irrelevant variance could confound the domain definition underlying the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.

Standard 3.18

For tests that have time limits, test development research should examine the degree to which scores include a speed component and evaluate the appropriateness of that component, given the domain the test is designed to measure.

Standard 3.19

The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained.

Comment: Because all people administering tests, including those in schools, industry, and clinics, need to follow test administration conditions carefully, it is essential that test administrators receive detailed instructions on test administration guidelines and procedures.

Standard 3.20

The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test's classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration instructions.

Comment: For example, in a personality inventory it may be intended that test takers give the first response that occurs to them. Such an expectation should be made clear in the inventory directions. As another example, in directions for interest or occupational inventories, it may be important to specify whether test takers are to mark the activities they would like ideally or whether they are to consider both their opportunity and their ability realistically.

The extent and nature of practice materials and directions depend on expected levels of knowledge among test takers. For example, in using a novel test format, it may be very important to provide the test taker a practice opportunity as part of the test administration. In some testing situations, it may be important for the instructions to address such matters as the effects that guessing and time limits have on test scores. If expansion or elaboration of the test instructions is permitted, the conditions under which this may be done should be stated clearly in the form of general rules and by giving representative examples. If no expansion or elaboration is to be permitted, this should be stated explicitly. Publishers should include guidance for dealing with typical questions from test takers. Users should be instructed how to deal with questions that may arise during the testing period.

Standard 3.21

If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified, and a rationale for permitting the different conditions should be documented.

Comment: In deciding whether the conditions of administration can vary, the test developer needs to consider and study the potential effects of varying conditions of administration. If conditions of administration vary from the conditions studied by the test developer or from those used in the development of norms, the comparability of the test scores may be weakened and the applicability of the norms can be questioned.

Standard 3.22

Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally.

Standard 3.23

The process for selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the

scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person(s) responsible for conducting the training session.

Standard 3.24

When scoring is done locally and requires scorer judgment, the test user is responsible for providing adequate training and instruction to the scorers and for examining scorer agreement and accuracy. The test developer should document the expected level of scorer agreement and accuracy.

Comment: A common practice of test developers is to provide examples of training materials (e.g., scoring rubrics, test takers' responses at each score level) and procedures when scoring is done locally and requires scorer judgment.

Standard 3.25

A test should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may lower the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.

Comment: Test developers need to consider a number of factors that may warrant the revision of a test, including outdated test content and language. If an older version of a test is used when a newer version has been published or made available, test users are responsible for providing evidence that the older version is as appropriate as the new version for that particular test use.

Standard 3.26

Tests should be labeled or advertised as "revised" only when they have been revised in significant ways. A phrase such as "with minor modification" should be used when the test has been modified in minor ways. The score scale should be adjusted to account for these modifications, and users should be informed of the adjustments made to the score scale.

Comment: It is the test developer's responsibility to determine whether revisions to a test would influence test score interpretations. If test score interpretations would be affected by the revisions, it would then be appropriate to label the test "revised." When tests are revised, the nature of the revisions and their implications on test score interpretations should be documented.

Standard 3.27

If a test or part of a test is intended for research use only and is not distributed for operational use, statements to this effect should be displayed prominently on all relevant test administration and interpretation materials that are provided to the test user.

Comment: This standard refers to tests that are intended for research use only and does not refer to standard test development functions that occur prior to the operational use of a test (e.g., field testing).

COMPARABILITY

Background

Test scores are reported on scales designed to assist score interpretation. Typically, scoring begins with responses to separate test items, which are often coded using 0 or 1 to represent wrong/right or negative/positive, but sometimes using numerical values to indicate finer response gradations. Then the item scores are combined, often by addition but sometimes by a more elaborate procedure, to obtain a *raw score.* Raw scores are determined, in part, by features of a test such as test length, choice of time limit, item difficulties, and the circumstances under which the test is administered. This makes raw scores difficult to interpret in the absence of further information. Interpretation and statistical analyses may be facilitated by converting raw scores into an entirely different set of values called *derived scores* or *scale scores.* The various scales used for reporting scores on college admissions tests, the standard scores often used to report results for intelligence scales or vocational interest and personality inventories, and the grade equivalents reported for achievement tests in the elementary grades are examples of scale scores. The process of developing such a score scale is called *scaling* a test. Scale scores may aid interpretation by indicating how a given score compares to those of other test takers, by enhancing the comparability of scores obtained using different forms of a test, or in other ways.

Another way of assisting score interpretation is to establish *standards* or *cut scores* that distinguish different score ranges. In some cases, a single cut score may define the boundary between passing and failing. In other cases, a series of cut scores may define distinct proficiency levels. Cut scores may be established for either raw or scale scores. Both scale scores and standards or cut scores can be central to the use and interpretation of test scores. For that reason, their defensibility is an important consideration in test validation. There is a close connection between standards or cut scores and certain scale scores. If the successive score ranges defined by a series of cut scores are relabeled, say 0, 1, 2, and so on, then a scale score has been created.

In addition to facilitating interpretations of a single test form considered in isolation, scale scores are often created to enhance comparability across different forms of the same test, across test formats or administration conditions, or even across tests designed to measure different constructs (e.g., related subtests in a battery). Equated scores from alternate forms of a test can often be interpreted more easily when expressed in scale score units rather than raw score units. Scaling may be used to place scores from different levels of an achievement test on a continuous scale and thereby facilitate inferences about growth or development. Scaling can also enhance the comparability of scores derived from tests in different areas, as in subtests within an aptitude, interest, or achievement battery.

Norm-Referenced and Criterion-Referenced Score Interpretations

Individual raw scores or scale scores are often referred to the distribution of scores for one or more comparison groups to draw useful inferences about an individual's performance. Test score interpretations based on such comparisons are said to be *norm-referenced.* Percentile rank norms, for example, indicate the standing of an individual or group within a defined population of individuals or groups. An example of such a comparison group might be fourth-grade students in the United States, tested in the last 2 months of a recent school year. Percentiles, averages, or other statistics for such reference groups are called *norms.* By showing

v the test score of a given examinee com-
es to those of others, norms assist in the
sification or description of examinees.

Other test score interpretations make no
ect reference to the performance of other
minees. These interpretations may take a
iety of forms; most are collectively referred
as *criterion-referenced* interpretations. Derived
ores supporting such interpretations may
dicate the likely proportion of correct
ponses on some larger domain of items, or
e probability of an examinee's answering
rticular sorts of items correctly. Other crite-
on-referenced interpretations may indicate
e likelihood that some psychopathology is
esent. Still other criterion-referenced inter-
etations indicate the probability that an
aminee's level of tested knowledge or skill
adequate to perform successfully in some
her setting; such probabilities may be sum-
arized in an expectancy table. Scale scores
support such criterion-referenced score
terpretations are often developed on the
asis of statistical analyses of the relationships
f test scores to other variables.

Some scale scores are developed primarily
support norm-referenced interpretations
nd others, criterion-referenced interpretations.
n practice, however, there is not always a sharp
istinction. Both criterion-referenced and
orm-referenced scales may be developed and
sed for the same test scores. Moreover, a
orm-referenced score scale originally devel-
ped, for example, to indicate performance
elative to some specific reference population
night, over time, also come to support crite-
ion-referenced interpretations. This could
appen as research and experience brought
ncreased understanding of the capabilities
implied by different scale score levels.
Conversely, results of an educational assess-
ment might be reported on a scale consisting
of several ordered proficiency levels, defined
by descriptions of the kinds of tasks students
at each level were able to perform. That would
be a criterion-referenced scale, but once the
distribution of scores over levels was reported,
say, for all eighth-grade students in a given
state, individual students' scores would also
convey information about their standing rela-
tive to that tested population.

Interpretations based on cut scores may
likewise be either criterion-referenced or
norm-referenced. If qualitatively different
descriptions are attached to successive score
ranges, a criterion-referenced interpretation is
supported. For example, the descriptions of
performance levels in some assessment task
scoring rubrics can enhance score interpreta-
tion by summarizing the capabilities that must
be demonstrated to merit a given score. In
other cases, criterion-referenced interpretations
may be based on empirically determined rela-
tionships between test scores and other vari-
ables. But when tests are used for selection, it
may be appropriate to rank-order examinees
according to their test performance and estab-
lish a cut score so as to select a prespecified
number or proportion of examinees from one
end of the distribution, if the selection use is
otherwise supported by relevant reliability
and validity evidence. In such cases, the cut
score interpretation is norm-referenced; the
labels *reject* or *fail* versus *accept* or *pass* are
determined solely by an examinee's standing
relative to others tested.

Criterion-referenced interpretations based
on cut scores are sometimes criticized on the
grounds that there is very rarely a sharp dis-
tinction of any kind between those just below
versus just above a cut score. A neuropsy-
chological test may be helpful in diagnosing
some particular impairment, for example, but
the probability that the impairment is pres-
ent is likely to increase continuously as a
function of the test score. Cut scores may
nonetheless aid in formulating rules for
reaching decisions on the basis of test per-
formance. It should be recognized, however,
that the probability of misclassification will
generally be relatively high for persons with
scores close to the cut points.

Norms

The validity of norm-referenced interpretations depends in part on the appropriateness of the reference group to which test scores are compared. Norms based on hospitalized patients, for example, might be inappropriate for some interpretations of nonhospitalized patients' scores. Thus, it is important that reference populations be carefully defined and clearly described. Validity of such interpretations also depends on the accuracy with which norms summarize the performance of the reference population. That population may be small enough that essentially the entire population can be tested (e.g., all pupils at a given grade level in a given district tested on the same occasion). Often, however, only a sample of examinees from the reference population is tested. It is then important that the norms be based on a technically sound, representative, scientific sample of sufficient size. Patients in a few hospitals in a small geographic region are unlikely to be representative of all patients in the United States, for example. Moreover, the appropriateness of norms based on a given sample may diminish over time. Thus, for tests that have been in use for a number of years, periodic review is generally required to assure the continued utility of norms. Renorming may be required to maintain the validity of norm-referenced test score interpretations.

More than one reference population may be appropriate for the same test. For example, achievement test performance might be interpreted by reference to local norms based on sampling from a particular school district, norms for a state or type of community, or national norms. For other tests, norms might be based on occupational or educational classifications. Descriptive statistics for all examinees who happen to be tested during a given period of time (sometimes called *user norms* or *program norms*) may be useful for some purposes, such as describing trends over time. But there must be sound reason to regard that group of test takers as an appropriate basis for such inferences. When there is a suitable rationale for using such a group, the descriptive statistics should be clearly characterized as being based on a sample of persons routinely tested as part of an ongoing program.

Comparability and Equating

Many test uses involve different versions of the same test, which yield scores that can be used interchangeably even though they are based on different sets of items. In testing programs that offer a choice of examination dates, for example, test security may be compromised if the same form is used repeatedly. Other testing applications may entail repeated measurements of the same individuals, perhaps to measure change in levels of psychological dysfunction, change in attitudes, or educational progress. In such contexts, reuse of the same set of test items may result in correlated errors of measurement and biased estimates of change. When distinct forms of a test are constructed to the same explicit content and statistical specifications and administered under identical conditions, they are referred to as *alternate forms* or sometimes *parallel* or *equivalent* forms. The process of placing scores from such alternate forms on a common scale is called *equating*. Equating is analogous to the calibration of different balances so that they all indicate the same weight for any given object. However, the equating process for test scores is more complex. It involves small statistical adjustments to account for minor differences in the difficulty and statistical properties of the alternate forms.

In theory, equating should provide accurate score conversions for any set of persons drawn from the examinee population for which the test is designed. Furthermore, the same score conversion should be appropriate regardless of the score interpretation or use intended. It is not possible to construct conversions with these ideal properties between scores on

hat measure different constructs; that materially in difficulty, reliability, time , or other conditions of administration; t are designed to different specifications. There is another assessment approach nay provide interchangeable scores based sponses to different items using different ods, not referred to as equating. This is se of *adaptive tests*. It has long been rec- ed that little is learned from examinees' nses to items that are much too easy or too difficult for them. Consequently, testing procedures use only a subset of vailable items with each examinee in to avoid boredom or frustration, or to en testing time. An adaptive test con- of a pool of items together with rules electing a subset of those items to be inistered to an individual examinee, and ocedure for placing different examinees' es on a common scale. The selection ccessive items is based in part on the inee's responses to previous items. The pool and item selection rules may be gned so that each examinee receives a esentative set of items, of appropriate culty. The selection rules generally re that an acceptable degree of precision tained before testing is terminated. At time, such tailored testing was limited ertain individually administered psy- logical tests. With advances in item onse theory (IRT) and in computer nology, however, adaptive testing is oming more sophisticated. With some ptive tests, it may happen that two minees rarely if ever respond to precisely same set of items. Moreover, two exam- s taking the same adaptive test may be n sets of items that differ markedly in iculty. Nevertheless, when certain statis- l and content conditions are met, test res produced by an adaptive testing sys- can function like scores from equated rnate forms.

Scaling to Achieve Comparability

The term *equating* is properly reserved only for score conversions derived for alternate forms of the same test. It is often useful, however, to compare scores from tests that cannot, in theory, be equated. For example, it may be desirable to interpret scores from a shortened (and hence less reliable) form of a test by first converting them to corresponding scores on the full-length version. For the evaluation of examinee growth over time, it may be desirable to develop scales that span a broad range of developmental or educational levels. Test revision often brings a need for some linkage between scores obtained using newer and older editions. International comparative studies or use with hearing-impaired examinees may require test forms in different languages. In still other cases, linkages or alignments may be created between tests measuring different constructs, perhaps comparing an aptitude with a form of behavior, or linking measures of achievement in several content areas. Scores from such tests may sometimes be aligned or presented in a concordance table to aid users in estimating relative performance on one test from performance on another.

Score conversions to facilitate such comparisons may be described using terms like linkage, calibration, concordance, projection, moderation, or anchoring. These weaker score linkages may be technically sound and may fully satisfy desired goals of comparability for one purpose or for one subgroup of examinees, but they cannot be assumed to be stable over time or invariant across multiple subgroups of the examinee population nor is there any assurance that scores obtained using different tests will be equally accurate. Thus, their use for other purposes or with other populations than originally intended may require additional research. For example, a score conversion that was accurate for a group of native speakers might systematically overpredict or underpredict the scores of a group of nonnative speakers.

Cut Scores

A critical step in the development and use of some tests is to establish one or more cut points dividing the score range to partition the distribution of scores into categories. These categories may be used just for descriptive purposes or may be used to distinguish among examinees for whom different programs are deemed desirable or different predictions are warranted. An employer may determine a cut score to screen potential employees or promote current employees; a school may use test scores to decide which of several alternative instructional programs would be most beneficial for a student; in granting a professional license, a state may specify a minimum passing score on a licensure test.

These examples differ in important respects, but all involve delineating categories of examinees on the basis of test scores. Such cut scores embody the rules according to which tests are used or interpreted. Thus, in some situations the validity of test interpretations may hinge on the cut scores. There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility. These examples serve only as illustrations.

The first example, that of an employer hiring all those who earn scores above a given level on an employment test, is most straightforward. Assuming that the employment test is valid for its intended use, average job performance would typically be expected to rise steadily, albeit slowly, with each increment in test score, at least for some range of scores surrounding the cut point. In such a case the designation of the particular value for the cut point may be largely determined by the number of persons to be hired or promoted. There is no sharp difference between those just below the cut point and those just above it, and the use of the cut score does not entail any criterion-referenced interpretation. This method of establishing a cut score may be subject to legal requirements with respect to the nature of the validity and reliability evidence needed to support the use of rank-order selections and the unavailability of effective alternative selection methods, if it has a disproportionate effect on one or more subgroups of employees or prospective employees.

In the second example, a school district might structure its courses in writing around three categories of needs. For children whose proficiency is least developed, instruction might be provided in small groups, with considerable individual attention to assist them in creating meaningful written stories grounded in their own experience. For children whose proficiency was further developed, more emphasis might be placed on systematic exploration of the stages of the writing process. Instruction for children at the highest proficiency level might emphasize mastery of specific writing genres or prose structures used in more formal writing. In an appropriate implementation of such a program, children could easily be transferred from one level to another if their original placement was in error or as their proficiency increased. Ideally, cut scores delineating categories in this application would be based on research demonstrating empirically that pupils in successive score ranges did most often benefit more from the respective treatments to which they were assigned than from the alternatives available. It would typically be found that between those score ranges in which one or another instructional treatment was clearly superior, there was an intermediate region in which neither treatment was clearly preferred. The cut score might be located somewhere in that intermediate region.

In the final example, that of a professional licensure examination, the cut score represents an informed judgment that those scoring below it are likely to make serious errors for want of the knowledge or skills tested. Little evidence apart from errors made on the test itself may document the need to deny the right to prac-

tice the profession. No test is perfect, of course, and regardless of the cut score chosen, some examinees with inadequate skills are likely to pass and some with adequate skills are likely to fail. The relative probabilities of such false positive and false negative errors will vary depending on the cut score chosen. A given probability of exposing the public to potential harm by issuing a license to an incompetent individual (false positive) must be weighed against some corresponding probability of denying a license to, and thereby disenfranchising, a qualified examinee (false negative). Changing the cut score to reduce either probability will increase the other, although both kinds of errors can be minimized through sound test design that anticipates the role of the cut score in test use and interpretation. Determining cut scores in such situations cannot be a purely technical matter, although empirical studies and statistical models can be of great value in informing the process.

Cut scores embody value judgments as well as technical and empirical considerations. Where the results of the standard-setting process have highly significant consequences, and especially where large numbers of examinees are involved, those responsible for establishing cut scores should be concerned that the process by which cut scores are determined be clearly documented and defensible. The qualifications of any judges involved in standard setting and the process by which they are selected are part of that documentation. Care must be taken to assure that judges understand what they are to do. The process must be such that well-qualified judges can apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and intentions. A sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were replicated.

Standard 4.1

Test documents should provide test users with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations.

Comment: All scales (raw score or derived) may be subject to misinterpretation. Sometimes scales are extrapolated beyond the range of available data or are interpolated without sufficient data points. Grade- and age-equivalent scores have been criticized in this regard, but percentile ranks and standard score scales are also subject to misinterpretation. If the nature or intended uses of a scale are novel, it is especially important that its uses, interpretations, and limitations be clearly described. Illustrations of appropriate versus inappropriate interpretations may be helpful, especially for types of scales or interpretations that may be unfamiliar to most users. This standard pertains to score scales intended for criterion-referenced as well as for norm-referenced interpretation.

Standard 4.2

The construction of scales used for reporting scores should be described clearly in test documents.

Comment: When scales, norms, or other interpretive systems are provided by the test developer, technical documentation should enable users to judge the quality and precision of the resulting derived scores. This standard pertains to score scales intended for criterion-referenced as well as for norm-referenced interpretation.

Standard 4.3

If there is sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly forewarned.

Comment: Test publishers and users can reduce misinterpretations of grade-equivalent scores, for example, by ensuring that such scores are accompanied by instructions that make clear that grade-equivalent scores do not represent a standard of growth per year or grade and that roughly 50% of the students tested in the standardization sample should by definition fall below grade level. As another example, a score scale point originally defined as the mean of some reference population should no longer be interpreted as representing average performance if the scale is held constant over time and the examinee population changes.

Standard 4.4

When raw scores are intended to be directly interpretable, their meanings, intended interpretations, and limitations should be described and justified in the same manner as is done for derived score scales.

Comment: In some cases the items in a test are a representative sample of a well-defined domain of items. The proportion correct on the test may then be interpreted as an estimate of the proportion of items in the domain that could be answered correctly. In other cases, different interpretations may be attached to scores above or below one or another cut score. Support should be offered for any such interpretations recommended by the test developer.

Standard 4.5

Norms, if used, should refer to clearly described populations. These populations should include individuals or groups to whom test users will ordinarily wish to compare their own examinees.

Comment: It is the responsibility of test developers to describe norms clearly and the responsibility of test users to employ norms appropriately. Users need to know the applicability of a test to different groups. Differentiated norms or summary information about differences between gender, ethnic, language, disability, grade, or age groups, for example, may be useful in some cases. The permissible uses of such differentiated norms and related information may be limited by law. Users also need to be made alert to situations in which norms are less appropriate for some groups or individuals than others. On an occupational interest inventory, for example, norms for persons actually engaged in an occupation may be inappropriate for interpreting the scores of persons not so engaged. As another example, the appropriateness of norms for personality inventories or relationship scales may differ depending upon an examinee's sexual orientation.

Standard 4.6

Reports of norming studies should include precise specification of the population that was sampled, sampling procedures and participation rates, any weighting of the sample, the dates of testing, and descriptive statistics. The information provided should be sufficient to enable users to judge the appropriateness of the norms for interpreting the scores of local examinees. Technical documentation should indicate the precision of the norms themselves.

Comment: Scientific sampling is important if norms are to be representative of intended populations. For example, schools already using a given published test and volunteering to participate in a norming study for that test should not be assumed to be representative of schools in general. In addition to sampling procedures, participation rates should be reported, and the method of calculating participation rates should be clearly described. Studies that are designed to be nationally representative often use weights so that the weighted sample better represents the nation than does the unweighted sample. When weights are used, it is important that the procedure for deriving the weights be described and that the demographic representa-

tion of both the weighted and the unweighted samples be given. If norming data are collected under conditions in which student motivation in completing the test is likely to differ from that expected during operational use, this should be clearly documented. Likewise, if the instructional histories of students in the norming sample differ systematically from those to be expected during operational test use, that fact should be noted. Norms based on samples cannot be perfectly precise. Even though the imprecision of norm-referenced interpretations due to imperfections in the norms themselves may be small compared to that due to measurement error, estimates of the precision of norms should be available in technical documentation. For example, standard errors based on the sample design might be presented. In some testing applications, norms based on all examinees tested over a given period of time may be useful for some purposes. Such norms should be clearly characterized as being based on a sample of persons routinely tested as part of an ongoing testing program.

Standard 4.7

If local examinee groups differ materially from the populations to which norms refer, a user who reports derived scores based on the published norms has the responsibility to describe such differences if they bear upon the interpretation of the reported scores.

Comment: In employment settings, the qualifications of local examinee groups may fluctuate depending on recruitment or referral procedures as well as market conditions. In such cases, appropriate test use and interpretation may not require documentation or cautions concerning departures from characteristics of the norming population.

Standard 4.8

When norms are used to characterize examinee groups, the statistics used to summarize each group's performance and the norms to which those statistics are referred should be clearly defined and should support the intended use or interpretation.

Comment: Group means are distributed differently from individual scores. For example, it is not possible to determine the percentile rank of a school's average test score if all that is known are the percentile ranks of each of that school's students. It may sometimes be useful to develop special norms for group means, but when the sizes of the groups differ materially or when some groups are much more heterogeneous than others, the construction and interpretation of group norms is problematical. One common and acceptable procedure is to report the percentile rank of the median group member, for example, the median percentile rank of the pupils tested in a given school.

Standard 4.9

When raw score or derived score scales are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained.

Comment: Criterion-referenced interpretations are score-based descriptions or inferences that do not take the form of comparisons to the test performance of other examinees. Examples include statements that some psychopathology is likely present, that a prospective employee possesses specific skills required in a given position, or that a child scoring above a certain score point can successfully apply a given set of skills. Such interpretations may refer to the absolute levels of test scores or to patterns of scores for an individual examinee. Whenever the test developer recommends such interpretations, the rationale and empirical basis should be clearly presented. Serious efforts should be made whenever possible to obtain independent evidence concerning the soundness of such score interpretations. Criterion-referenced and norm-referenced scales are not mutually exclusive. Given adequate supporting data, scores may be interpreted by both approaches, not necessarily just one or the other.

Standard 4.10

A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably. In some cases, direct evidence of score equivalence may be provided. In other cases, evidence may come from a demonstration that the theoretical assumptions underlying procedures for establishing score comparability have been sufficiently satisfied. The specific rationale and the evidence required will depend in part on the intended uses for which score equivalence is claimed.

Comment: Support should be provided for any assertion that scores obtained using different items or testing materials, or different testing procedures, are interchangeable for some purpose. This standard applies, for example, to alternate forms of a paper-and-pencil test or to alternate sets of items taken by different examinees in computerized adaptive testing. It also applies to test forms administered in different formats (e.g., paper-and-pencil and computerized tests) or test forms designed for individual versus group administration. Score equivalence is easiest to establish when different forms are constructed following identical procedures and then equated statistically. When that is not possible, for example, in cases where different test formats are used, additional evidence may be required to establish the requisite degree of score equivalence for the intended context and purpose. When recommended inferences or actions are based solely on classifications of examinees into one of two or more categories, the rationale and evidence should address consistency of classification. If the only score reported and used is a pass-fail decision for example, then the form-to-form equivalence of measurements for examinees far above or far below the cut score is of no concern. Some testing accommodations may only affect the dependence of test scores on capabilities irrelevant to the construct the test is intended to measure. Use of a large-print edition, for example, assures that performance does not depend on the ability to perceive standard-size print. In such cases, relatively modest studies or professional judgment may be sufficient to support claims of score equivalence.

Standard 4.11

When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions.

Comment: The fundamental concern is to show that equated scores measure essentially the same construct, with very similar levels of reliability and conditional standard errors of measurement. Technical information should include the design of equating studies, the statistical methods used, the size and relevant characteristics of examinee samples used in equating studies, and the characteristics of any anchor tests or linking items. Standard errors of equating functions should be estimated and reported whenever possible. Sample sizes permitting, it may be informative to determine equating functions independently for identifiable subgroups of examinees. It may also be informative to use two anchor forms and to conduct the equating using each of the anchors. In some cases, equating functions may be determined independently using different statistical methods. The correspondence of separate functions obtained by such methods can lend support to the adequacy of the equating results. An substantial disparities found by such method

hould be resolved or reported. To be most seful, equating error should be presented in nits of the reported score scale. For testing rograms with cut scores, equating error near he cut score is of primary importance. The legree of scrutiny of equating functions should e commensurate with the extent of test use nticipated and the importance of the deci- ions the test scores are intended to inform.

Standard 4.12

n equating studies that rely on the statisti- al equivalence of examinee groups receiving different forms, methods of assuring such equivalence should be described in detail.

Comment: Certain equating designs rely on the andom equivalence of groups receiving different orms. Often, one way to assure such equivalence s to systematically mix different test forms and hen distribute them in a random fashion so that roughly equal numbers of examinees in each group tested receive each form.

Standard 4.13

In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-based and classical equating studies, the representativeness and psychometric characteristics of anchor items should be presented.

Comment: Tests or test forms may be linked via common items embedded within each of them, or a common test administered together with each of them. These common items or tests are referred to as linking items, anchor items, or anchor tests. With such methods, the quality of the resulting equating depends strongly on the adequacy of the anchor tests or items used.

Standard 4.14

When score conversions or comparison procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those conversions or comparisons should be clearly described.

Comment: Various score conversions or concordance tables have been constructed relating tests at different levels of difficulty, relating earlier to revised forms of published tests, creating score concordances between different tests of similar or different constructs, or for other purposes. Such conversions are often useful, but they may also be subject to misinterpretation. The limitations of such conversions should be clearly described.

Standard 4.15

When additional test forms are created by taking a subset of the items in an existing test form or by rearranging its items and there is sound reason to believe that scores on these forms may be influenced by item context effects, evidence should be provided that there is no undue distortion of norms for the different versions or of score linkages between them.

Comment: Some tests and test batteries are published in both a full-length version and a survey or short version. In other cases, multiple versions of a single test form may be created by rearranging its items. It should not be assumed that performance data derived from the administration of items as part of the initial version can be used to approximate norms or construct conversion tables for alternative intact tests. Due caution is required in cases where context effects are likely, including speeded tests, long tests where fatigue may be a factor, and so on. In many cases, adequate psychometric data may only be obtainable from independent administrations of the alternate forms.

Standard 4.16

If test specifications are changed from one version of a test to a subsequent version, such changes should be identified in the test manual, and an indication should be given that converted scores for the two versions may not be strictly equivalent. When substantial changes in test specifications occur, either scores should be reported on a new scale or a clear statement should be provided to alert users that the scores are not directly comparable with those on earlier versions of the test.

Comment: Major shifts sometimes occur in the specifications of tests that are used for substantial periods of time. Often such changes take advantage of improvements in item types or of shifts in content that have been shown to improve validity and, therefore, are highly desirable. It is important to recognize, however, that such shifts will result in scores that cannot be made strictly interchangeable with scores on an earlier form of the test.

Standard 4.17

Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported.

Comment: In some testing programs, items are introduced into and retired from item pools on an ongoing basis. In other cases, the items in successive test forms may overlap very little, or not at all. In either case, if a fixed scale is used for reporting, it is important to assure that the meaning of the scaled scores does not change over time.

Standard 4.18

If a publisher provides norms for use in test score interpretation, then so long as the test remains in print, it is the publisher's responsibility to assure that the test is renormed with sufficient frequency to permit continued accurate and appropriate score interpretations.

Comment: Test publishers should assure that up-to-date norms are readily available, but it remains the test user's responsibility to avoid inappropriate use of norms that are out of date and to strive to assure accurate and appropriate test interpretations.

Standard 4.19

When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.

Comment: Cut scores may be established to select a specified number of examinees (e.g., to fill existing vacancies), in which case little further documentation may be needed concerning the specific question of how the cut scores are established, though attention should be paid to legal requirements that may apply. In other cases, however, cut scores may be used to classify examinees into distinct categories (e.g., diagnostic categories, or passing versus failing) for which there are no preestablished quotas. In these cases, the standard-setting method must be clearly documented. Ideally, the role of cut scores in test use and interpretation is taken into account during test design. Adequate precision in regions of score scales where cut points are established is prerequisite to reliable classification of examinees into categories. If standard setting employs data on the score distributions for criterion groups or on the relation of test scores to one or more criterion variables, those data should be summarized in technical documentation. If a judgmental standard-setting process is followed, the method employed should be clearly described, and the precise nature of the judgments called for should be presented, whether those are judgments of persons, of item or test performances, or of other criterion performances predicted by test scores. Documentation should also include the selection and qualification of judges, training provided, any feedback to judges concerning the implications of their provisional judgments,

and any opportunities for judges to confer with one another. Where applicable, variability over judges should be reported. Whenever feasible, an estimate should be provided of the amount of variation in cut scores that might be expected if the standard-setting procedure were replicated.

Standard 4.20

When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.

Comment: In employment settings, although it is important to establish that test scores are related to job performance, the precise relation of test and criterion may have little bearing on the choice of a cut score. However, in contexts where distinct interpretations are applied to different score categories, the empirical relation of test to criterion assumes greater importance. Cut scores used in interpreting diagnostic tests may be established on the basis of empirically determined score distributions for criterion groups. With achievement or proficiency tests, such as those used in licensure, suitable criterion groups (e.g., successful versus unsuccessful practitioners) are often unavailable. Nonetheless, it is highly desirable, when appropriate and feasible, to investigate the relation between test scores and performance in relevant practical settings. Note that a carefully designed and implemented procedure based solely on judgments of content relevance and item difficulty may be preferable to an empirical study with an inadequate criterion measure or other deficiencies. Professional judgment is required to determine an appropriate standard-setting approach (or combination of approaches) in any given situation. In general, one would not expect to find a sharp difference in levels of the criterion variable between those just below versus just above the cut score, but evidence should be provided where feasible of a relationship between test and criterion performance over a score interval that includes or approaches the cut score.

Standard 4.21

When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

Comment: Cut scores are sometimes based on judgments about the adequacy of item or test performances (e.g., essay responses to a writing prompt) or performance levels (e.g., the level that would characterize a borderline examinee). The procedures used to elicit such judgments should result in reasonable, defensible standards that accurately reflect the judges' values and intentions. Reaching such judgments may be most straightforward when judges are asked to consider kinds of performances with which they are familiar and for which they have formed clear conceptions of adequacy or quality. When the responses elicited by a test neither sample nor closely simulate the use of tested knowledge or skills in the actual criterion domain, judges are not likely to approach the task with such clear understandings. Special care must then be taken to assure that judges have a sound basis for making the judgments requested. Thorough familiarity with descriptions of different proficiency categories, practice in judging task difficulty with feedback on accuracy, the experience of actually taking a form of the test, feedback on the failure rates entailed by provisional standards, and other forms of information may be beneficial in helping judges to reach sound and principled decisions.

Background

The usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions. When directions to examinees, testing conditions, and scoring procedures follow the same detailed procedures, the test is said to be standardized. Without such standardization, the accuracy and comparability of score interpretations would be reduced. For tests designed to assess the examinee's knowledge, skills, or abilities, standardization helps to ensure that all examinees have the same opportunity to demonstrate their competencies. Maintaining test security also helps to ensure that no one has an unfair advantage.

Occasionally, however, situations arise in which modifications of standardized procedures may be advisable or legally mandated. Persons of different backgrounds, ages, or familiarity with testing may need nonstandard modes of test administration or a more comprehensive orientation to the testing process, in order that all test takers can come to the same understanding of the task. Standardized modes of presenting information or of responding may not be suitable for specific individuals, such as persons with some kinds of disability, or persons with limited proficiency in the language of the test, so that accommodations may be needed (see chapters 9 and 10). Large-scale testing programs generally have established specific procedures to be used in considering and granting accommodations. Some test users feel that any accommodation not specifically required by law could lead to a charge of unfair treatment and discrimination. Although accommodations are made with the intent of maintaining score comparability, the extent to which that is possible may not be known. Comparability of scores may be compromised, and the test may then not measure the same constructs for all test takers.

Tests and assessments differ in their degree of standardization. In many instances different examinees are given not the same test form, but equivalent forms that have been shown to yield comparable scores. Some assessments permit examinees to choose which tasks to perform or which pieces of their work are to be evaluated. A degree of standardization can be maintained by specifying the conditions of the choice and the criteria of evaluation of the products. When an assessment permits a certain kind of collaboration, the limits of that collaboration can be specified. With some assessments, test administrators may be expected to tailor their instructions to help assure that all examinees understand what is expected of them. In all such cases, the goal remains the same: to provide accurate and comparable measurement for everyone, and unfair advantage to no one. The degree of standardization is dictated by that goal, and by the intended use of the test.

Standardized directions to test takers help to ensure that all test takers understand the mechanics of test taking. Directions generally inform test takers how to make their responses, what kind of help they may legitimately be given if they do not understand the question or task, how they can correct inadvertent responses, and the nature of any time constraints. General advice is sometimes given about omitting item responses. Many tests, including computer-administered tests, require special equipment. Practice exercises are often presented in such cases to ensure that the test taker understands how to operate the equipment. The principle of standardization includes orienting test takers to materials with which they may not be familiar. Some equipment may be provided at the testing site, such as shop tools or balances. Opportunity for test takers to practice with the equipment will often be appropriate, unless using the equipment is the purpose of the test.

computer, with test responses made by keyboard, computer mouse, or similar device. Although many test takers are accustomed to computers, some are not and may need some brief explanation. Even those test takers who use computers will need to know about some details. Special issues arise in managing the testing environment, such as the arrangement of illumination so that light sources do not reflect on the computer screen, possibly interfering with display legibility. Maintaining a quiet environment can be challenging when candidates are tested separately, starting at different times and finishing at different times from neighboring test takers. Those who administer computer-based tests require training in the hardware and software used for the test, so that they can deal with problems that may arise in human-computer interactions.

Standardized scoring procedures help to ensure accurate scoring and reporting, which are essential in all circumstances. When scoring is done by machine, the accuracy of the machine is at issue, including any scoring algorithm. When scoring is done by human judges, scorers require careful training. Regular monitoring can also help to ensure that every test protocol is scored according to the same standardized criteria and that the criteria do not change as the test scorers progress through the submitted test responses.

Test scores, per se, are not readily interpreted without other information, such as norms or standards, indications of measurement error, and descriptions of test content. Just as a temperature of 50° in January is warm for Minnesota and cool for Florida, a test score of 50 is not meaningful without some context. When the scores are to be reported to persons who are not technical specialists, interpretive material can be provided that is readily understandable to those receiving the report. Often, the test user provides an interpretation of the results for the test taker, suggesting the limitations of the results and the relationship of any reported scores to other information. Scores on some tests are not designed to be released to test takers; only broad test interpretations, or dichotomous classifications, such as pass/fail, are intended to be reported.

Interpretations of test results are sometimes prepared by computer systems. Such interpretations are generally based on a combination of empirical data and expert judgment and experience. In some professional applications of individualized testing, the computer-prepared interpretations are communicated by a professional, possibly with modifications for special circumstances. Such test interpretations require validation. Consistency with interpretations provided by nonalgorithmic approaches is clearly a concern.

In some large-scale assessments, the primary target of assessment is not the individual test taker but is a larger unit, such as a school district or an industrial plant. Often, different test takers are given different sets of items, following a carefully balanced matrix sampling plan, to broaden the range of information that can be obtained in a reasonable time period. The results acquire meaning when aggregated over many individuals taking different samples of items. Such assessments may not furnish enough information to support even minimally valid, reliable scores for individuals, as each individual may take only an incomplete test.

Some further issues of administration and scoring are discussed in chapter 3, "Test Development and Revision."

Standard 5.1

Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation or a test taker's disability dictates that an exception should be made.

Comment: Specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment should be strictly observed. In general, the same procedures should be followed as were used when obtaining the data for scaling and norming the test scores. A test taker with a disabling condition may require special accommodation. Other special circumstances may require some flexibility in administration. Judgments of the suitability of adjustments should be tempered by the consideration that departures from standard procedures may jeopardize the validity of the test score interpretations.

Standard 5.2

Modifications or disruptions of standardized test administration procedures or scoring should be documented.

Comment: Information about the nature of modifications of administration should be maintained in secure data files, so that research studies or case reviews based on test records can take this into account. This includes not only special accommodations for particular test takers, but also disruptions in the testing environment that may affect all test takers in the testing session. A researcher may wish to use only the records based on standardized administration. In other cases, research studies may depend on such information to form groups of respondents. Test users or test sponsors should establish policies concerning who keeps the files and who may have access to the files. Whether the information about modifications is reported to users of test data, such as admissions officers, depends on different considerations (see chapters 8 and 10). If such reports are made, certain cautions may be appropriate.

Standard 5.3

When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.

Comment: When large-scale testing programs have established strict procedures to be followed, administrators should not depart from these procedures.

Standard 5.4

The testing environment should furnish reasonable comfort with minimal distractions.

Comment: Noise, disruption in the testing area, extremes of temperature, poor lighting, inadequate work space, illegible materials, and so forth are among the conditions that should be avoided in testing situations. The testing site should be readily accessible. Testing sessions should be monitored where appropriate to assist the test taker when a need arises and to maintain proper administrative procedures. In general, the testing conditions should be equivalent to those that prevailed when norms and other interpretative data were obtained.

Standard 5.5

Instructions to test takers should clearly indicate how to make responses. Instructions should also be given in the use of any equipment likely to be unfamiliar to test takers. Opportunity to practice responding should be given when equipment is involved, unless use of the equipment is being assessed.

Comment: When electronic calculators are provided for use, examinees may need practice in using the calculator. Examinees may need practice responding with unfamiliar tasks, such as a numeric grid, which is sometimes used with mathematics performance items. In computer-administered tests, the method of responding may be unfamiliar to some test takers. Where possible, the practice responses should be monitored to ensure that the test taker is making acceptable responses. In some performance tests that involve tools or equipment, instructions may be needed for unfamiliar tools, unless accommodating to unfamiliar tools is part of what is being assessed. If a test taker is unable to use the equipment or make the responses, it may be appropriate to consider alternative testing modes.

Standard 5.6

Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.

Comment: In large-scale testing programs where the results may be viewed as having important consequences, efforts to assure score integrity should include, when appropriate and practicable, stipulating requirements for identification, constructing seating charts, assigning test takers to seats, requiring appropriate space between seats, and providing continuous monitoring of the testing process. Test developers should design test materials and procedures to minimize the possibility of cheating. Test administrators should note and report any significant instances of testing irregularity. A local change in the date or time of testing may offer an opportunity for fraud. In general, steps should be taken to minimize the possibility of breaches in test security. In any evaluation of work products (e.g., portfolios) steps should be taken to ensure that the product represents the candidate's own work, and that the amount and kind of assistance provided should be consistent with the intent of the assessment. Ancillary documentation, such as the date when the work was done, may be useful.

Standard 5.7

Test users have the responsibility of protecting the security of test materials at all times.

Comment: Those who have test materials under their control should, with due consideration of ethical and legal requirements, take all steps necessary to assure that only individuals with a legitimate need for access to test materials are able to obtain such access before the test administration, and afterwards as well, if any part of the test will be reused at a later time. Test users must balance test security with the rights of all test takers and test users. When sensitive test documents are challenged, it may be appropriate to employ an independent third party, using a closely supervised secure procedure to conduct a review of the relevant materials. Such secure procedures are usually preferable to placing tests, manuals, and an examinee's test responses in the public record.

Standard 5.8

Test scoring services should document the procedures that were followed to assure accuracy of scoring. The frequency of scoring errors should be monitored and reported to users of the service on reasonable request. Any systematic source of scoring errors should be corrected.

Comment: Clerical and mechanical errors should be examined. Scoring errors should be minimized and, when they are found, steps should be taken promptly to minimize their recurrence.

Standard 5.9

When test scoring involves human judgment, scoring rubrics should specify criteria for scoring. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented.

Comment: Human scorers may be provided with scoring rubrics listing acceptable alternative responses, as well as general criteria. Consistency of scoring is often checked by rescoring randomly selected test responses and by rescoring some responses from earlier administrations. Periodic checks of the statistical properties (e.g., means, standard deviations) of scores assigned by individual scorers during a scoring session can provide feedback for the scorers, helping them to maintain scoring standards. Lack of consistent scoring may call for retraining or dismissing some scorers or for reexamining the scoring rubrics.

Standard 5.10

When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

Comment: Test users should consult the interpretive material prepared by the test developer or publisher and should revise or supplement the material as necessary to present the local and individual results accurately and clearly. Score precision might be depicted by error bands, or likely score ranges, showing the standard error of measurement.

Standard 5.11

When computer-prepared interpretations of test response protocols are reported, the sources, rationale, and empirical basis for these interpretations should be available, and their limitations should be described.

Comment: Whereas computer-prepared interpretations may be based on expert judgment, the interpretations are of necessity based on accumulated experience and may not be able to take into consideration the context of the individual's circumstances. Computer-prepared interpretations should be used with care in diagnostic settings, because they may not take into account other information about the individual test taker, such as age, gender, education, prior employment, and medical history, that provide context for test results.

Standard 5.12

When group-level information is obtained by aggregating the results of partial tests taken by individuals, validity and reliability should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established.

Comment: Large-scale assessments often achieve efficiency by "matrix sampling" of the content domain by asking different test takers different questions. The testing then requires less time from each test taker, while the aggregation of individual results provides for domain coverage that can be adequate for meaningful group- or program-level interpretations, such as schools, or grade levels within a locality or particular subject-matter areas. Because the individual receives only an incomplete test, an individual score would have limited meaning. If individual scores are provided, comparisons between scores obtained by different individuals are based on responses to items that may cover different material. Some degree of calibration among incomplete tests can sometimes be made. Such calibration is essential to the comparisons of individual scores.

Standard 5.13

Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores.

Comment: Care is always needed when communicating the scores of identified test takers, regardless of the form of communication. Face-to-face communication, as well as telephone and written communication present well-known problems. Transmission by electronic media, including computer networks and facsimile, presents modern challenges to confidentiality.

Standard 5.14

When a material error is found in test scores or other important information released by a testing organization or other institution, a corrected score report should be distributed as soon as practicable to all known recipients who might otherwise use the erroneous scores as a basis for decision making. The corrected report should be labeled as such.

Comment: A material error is one that could change the interpretation of the test score. Innocuous typographical errors would be excluded. Timeliness is essential for decisions that will be made soon after the test scores are received.

Standard 5.15

When test data about a person are retained, both the test protocol and any written report should also be preserved in some form. Test users should adhere to the policies and record-keeping practice of their professional organizations.

Comment: The protocol may be needed to respond to a possible challenge from a test taker. The protocol would ordinarily be accompanied by testing materials and test scores. Retention of more detailed records of responses would depend on circumstances and should be covered in a retention policy (see the following standard). Record keeping may be subject to legal and professional requirements. Policy for the release of any test information for other than research purposes is discussed in chapter 8.

Standard 5.16

Organizations that maintain test scores on individuals in data files or in an individual's records should develop a clear set of policy guidelines on the duration of retention of an individual's records, and on the availability, and use over time, of such data.

Comment: In some instances, test scores become obsolete over time, no longer reflecting the current state of the test taker. Outdated scores should generally not be used or made available, except for research purposes. In other cases, test scores obtained in past years can be useful as, for example, in longitudinal assessment. The key issue is the valid use of the information. Score retention and disclosure may be subject to legal and professional requirements.

FOR TESTS

Background

The provision of supporting documents for tests is the primary means by which test developers, publishers, and distributors communicate with test users. These documents are evaluated on the basis of their completeness, accuracy, currency, and clarity and should be available to qualified individuals as appropriate. A test's documentation typically specifies the nature of the test; its intended use; the processes involved in the test's development; technical information related to scoring, interpretation, and evidence of validity and reliability; scaling and norming if appropriate to the instrument; and guidelines for test administration and interpretation. The objective of the documentation is to provide test users with the information needed to make sound judgments about the nature and quality of the test, the resulting scores, and the interpretations based on the test scores. The information may be reported in documents such as test manuals, technical manuals, user's guides, specimen sets, examination kits, directions for test administrators and scorers, or preview materials for test takers.

Test documentation is most effective if it communicates information to multiple user groups. To accommodate the breadth of training of professionals who use tests, separate documents or sections of documents may be written for identifiable categories of users such as practitioners, consultants, administrators, researchers, and educators. For example, the test user who administers the tests and interprets the results needs interpretive information or guidelines. On the other hand, those who are responsible for selecting tests need to be able to judge the technical adequacy of the test. Therefore, some combination of technical manuals, user's guides, test manuals, test supplements, examination kits, or specimen sets ordinarily is published to provide a potential test user or test reviewer with sufficient information to evaluate the appropriateness and technical adequacy of the test. The types of information presented in these documents typically include a description of the intended test-taking population, stated purpose of the test, test specifications, item formats, scoring procedures, and the test development process. Technical data, such as psychometric indices of the items, reliability and validity evidence, normative data, and cut scores or configural rules including those for computer-generated interpretations of test scores also are summarized.

An essential feature of the documentation for every test is a discussion of the known appropriate and inappropriate uses and interpretations of the test scores. The inclusion of illustrations of score interpretations, as they relate to the test developer's intended applications, also will help users make accurate inferences on the basis of the test scores. When possible, illustrations of improper test uses and inappropriate test score interpretations will help guard against the misuse of the test.

Test documents need to include enough information to allow test users and reviewers to determine the appropriateness of the test for its intended purposes. References to other materials that provide more details about research by the publisher or independent investigators should be cited and should be readily obtainable by the test user or reviewer. This supplemental material can be provided in any of a variety of published or unpublished forms; when demand is likely to be low, it may be maintained in archival form, including electronic storage. Test documentation is useful for all test instruments, including those that are developed exclusively for use within a single organization.

In addition to technical documentation, descriptive materials are needed in some settings to inform examinees and other interested parties about the nature and content of the test. The amount and type of information will depend on the particular test and application. For example, in situations requiring informed consent, information should be sufficient to develop a reasoned judgment. Such information should be phrased in nontechnical language and should be as inclusive as is consistent with the use of the test scores. The materials may include a general description and rationale for the test; sample items or complete sample tests; and information about conditions of test administration, confidentiality, and retention of test results. For some applications, however, the true nature and purpose of a test are purposely hidden or disguised to prevent faking or response bias. In these instances, examinees may be motivated to reveal more or less of the characteristics intended to be assessed. Under these circumstances, hiding or disguising the true nature or purpose of the test is acceptable provided this action is consistent with legal principles and ethical standards.

This chapter provides general standards for the preparation and publication of test documentation. The other chapters contain specific standards that will be useful to test developers, publishers, and distributors in the preparation of materials to be included in a test's documentation.

Standard 6.1

Test documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use.

Comment: The test developer or publisher should judge carefully which information should be included in first editions of the test manual, technical manual, or user's guides and which information can be provided in supplements. For low-volume, unpublished tests, the documentation may be relatively brief. When the developer is also the user, documentation and summaries are still necessary.

Standard 6.2

Test documents should be complete, accurate, and clearly written so that the intended reader can readily understand the content.

Comment: Test documents should provide sufficient detail to permit reviewers and researchers to judge or replicate important analyses published in the test manual. For example, reporting correlation matrices in the test document may allow the test user to judge the data upon which decisions and conclusions were based, or describing in detail the sample and the nature of any factor analyses that were conducted will allow the test user to replicate reported studies.

Standard 6.3

The rationale for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Comment: Test publishers make every effort to caution test users against known misuses of tests. However, test publishers are not required to anticipate all possible misuses of a test. If publishers do know of persistent test misuse by a test user, extraordinary educational efforts may be appropriate.

Standard 6.4

The population for whom the test is intended and the test specifications should be documented. If applicable, the item pool and scale development procedures should be described in the relevant test manuals. If normative data are provided, the norming population should be described in terms of relevant demographic variables, and the year(s) in which the data were collected should be reported.

Comment: Known limitations of a test for certain populations also should be clearly delineated in the test documents. In addition, if the test is available in more than one language, test documents should provide information on the translation or adaptation procedures, on the demographics of each norming sample, and on score interpretation issues for each language into which the test has been translated.

Standard 6.5

When statistical descriptions and analyses that provide evidence of the reliability of scores and the validity of their recommended interpretations are available, the information should be included in the test's documentation. When relevant for test interpretation, test documents ordinarily should include item level information, cut scores and configural rules, information about raw scores and derived scores, normative data, the standard errors of measurement, and a description of the procedures used to equate multiple forms.

Standard 6.6

When a test relates to a course of training or study, a curriculum, a textbook, or packaged instruction, the documentation should include an identification and description of the course or instructional materials and should indicate the year in which these materials were prepared.

Standard 6.7

Test documents should specify qualifications that are required to administer a test and to interpret the test scores accurately.

Comment: Statements of user qualifications need to specify the training, certification, competencies, or experience needed to have access to a test.

Standard 6.8

If a test is designed to be scored or interpreted by test takers, the publisher and test developer should provide evidence that the test can be accurately scored or interpreted by the test takers. Tests that are designed to be scored and interpreted by the test taker should be accompanied by interpretive materials that assist the individual in understanding the test scores and that are written in language that the test taker can understand.

Standard 6.9

Test documents should cite a representative set of the available studies pertaining to general and specific uses of the test.

Comment: Summaries of cited studies—excluding published works, dissertations, or proprietary documents—should be made available on request to test users and researchers by the publisher.

Standard 6.10

Interpretive materials for tests, that include case studies, should provide examples illustrating the diversity of prospective test takers.

Comment: For some instruments, the presentation of case studies that are intended to

scores and profiles also will be appropriate for inclusion in the test documentation. For example, case studies might cite as appropriate examples of women and men of different ages; individuals differing in sexual orientation; persons representing various ethnic, cultural, or racial groups; and individuals with special needs. The inclusion of examples illustrating the diversity of prospective test takers is not intended to promote interpretation of test scores in a manner inconsistent with legal requirements that may restrict certain practices in some contexts, such as employee selection.

Standard 6.11

If a test is designed so that more than one method can be used for administration or for recording responses—such as marking responses in a test booklet, on a separate answer sheet, or on a computer keyboard—then the manual should clearly document the extent to which scores arising from these methods are interchangeable. If the results are not interchangeable, this fact should be reported, and guidance should be given for the interpretation of scores obtained under the various conditions or methods of administration.

Standard 6.12

Publishers and scoring services that offer computer-generated interpretations of test scores should provide a summary of the evidence supporting the interpretations given.

Comment: The test user should be informed of any cut scores or configural rules necessary for understanding computer-generated score interpretations. A description of both the samples used to derive cut scores or configural rules and the methods used to derive the cut scores should be provided. When proprietary interests result in the withholding of cut scores or configural rules, the owners of the intellectual property are responsible for documenting evidence in support of the validity of computer-generated score interpretations. Such evidence might be provided, for example, by reporting the finding of an independent review of the algorithms by qualified professionals.

Standard 6.13

When substantial changes are made to a test, the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or cautions.

Standard 6.14

Every test form and supporting document should carry a copyright date or publication date.

Comment: During the operational life of a test, new or revised test forms may be published, and manuals and other materials may be added or revised. Users and potential users are entitled to know the publication dates of various documents that include test norms. Communication among researchers is hampered when the particular test documents used in experimental studies are ambiguously referenced in research reports.

Standard 6.15

Test developers, publishers, and distributors should provide general information for test users and researchers who may be required to determine the appropriateness of an intended test use in a specific context. When a particular test use cannot be justified, the response to an inquiry from a prospective test user should indicate this fact clearly. General information also should be provided for test takers and legal guardians who must provide consent prior to a test's administration.

PART II

Fairness in Testing