

# ETS STANDARDS

---

*for*

*Quality*

*and*

*Fairness*

2002

# ETS STANDARDS

---

*for*  
*Quality*

*and*  
*Fairness*



---

Educational Testing Service  
Princeton, NJ



Educational Testing Service is the world's premier educational testing organization and a leader in educational measurement research. A private nonprofit company, it is dedicated to serving the needs of individuals, educational institutions and agencies, and governmental bodies in 181 countries. ETS develops and annually administers more than 11 million tests worldwide on behalf of clients in education, government and business. For more information, access the ETS Web site at [www.ets.org](http://www.ets.org).

Copyright © 2002 by Educational Testing Service. All rights reserved.

Educational Testing Service is an Affirmative Action/Equal Opportunity Employer.

EDUCATIONAL TESTING SERVICE, ETS, and the ETS logos are registered trademarks of Educational Testing Service.

---

Published by Educational Testing Service  
Rosedale Road  
Princeton, NJ 08541

# Table of Contents

<b>Preface</b> .....	v
<b>Introduction</b> .....	1
<b>Overview</b> .....	3
<b>Audit Requirements</b> .....	5
<b>Chapter 1: Developmental Procedures</b> .....	9
Purpose .....	9
Standards 1.1–1.5 .....	9
<b>Chapter 2: Suitability for Use</b> .....	11
Purpose .....	11
Standards 2.1–2.6 .....	11
<b>Chapter 3: Customer Service</b> .....	13
Purpose .....	13
Standards 3.1–3.8 .....	13
<b>Chapter 4: Fairness</b> .....	17
Purpose .....	17
Standards 4.1–4.8 .....	18
<b>Chapter 5: Uses and Protection of Information</b> .....	23
Purpose .....	23
Standards 5.1–5.8 .....	23
<b>Chapter 6: Validity</b> .....	27
Purpose .....	27
Standards 6.1–6.8 .....	27
<b>Chapter 7: Assessment Development</b> .....	33
Purpose .....	33
Standards 7.1–7.8 .....	33
<b>Chapter 8: Reliability</b> .....	39
Purpose .....	39
Standards 8.1–8.6 .....	39
<b>Chapter 9: Cut Scores, Scaling, and Equating</b> .....	45
Purpose .....	45
Standards 9.1–9.6 .....	45

## Table of Contents

<b>Chapter 10: Assessment Administration</b> .....	49
Purpose .....	49
Standards <b>10.1–10.5</b> .....	49
<b>Chapter 11: Reporting Assessment Results</b> .....	53
Purpose .....	53
Standards <b>11.1–11.6</b> .....	53
<b>Chapter 12: Assessment Use</b> .....	57
Purpose .....	57
Standards <b>12.1–12.6</b> .....	57
<b>Chapter 13: Test Takers’ Rights and Responsibilities</b> .....	61
Purpose .....	61
Standards <b>13.1–13.5</b> .....	61
<b>Glossary</b> .....	65

# Preface

Our mission at Educational Testing Service is to help advance quality and equity in education by providing fair and valid assessments, research, and related services. Central to this mission are the *ETS Standards for Quality and Fairness*, which were adopted as corporate policy by the ETS Board of Trustees in 1981, and which continue to serve as the benchmark of excellence for all ETS products and services.

The Standards were revised in 2000, following the revision of the *Standards for Educational and Psychological Testing*, published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The two sets of standards are comparable, differing in areas that reflect the specific nature of ETS product and service offerings.

The ETS Standards are used by ETS staff on a daily basis, and successful and comprehensive implementation is checked through a rigorous audit program. The ETS Board of Trustees oversees the audit program results. In order for ETS to continue to be a global leader in providing fair and valid assessments, research, and related products and services, we must continue to hold ourselves accountable for meeting the highest levels of quality and fairness in our products.

Our products are aimed at helping individuals, parents, teachers, educational institutions, businesses, governments, countries, states, school districts, measurement specialists, and researchers. We are committed to ensuring integrity in all that we do and to providing outstanding customer service that meets or exceeds the expectations of our clients. All of our constituents count on the soundness and high quality of our tests and other products.

As I have said before, any testing program is only as good as the weakest link in the test development process. Ensuring the quality and fairness of tests is essential to test development; by using the benchmark of the ETS Standards, testing programs must meet standards for fairness, reliability, validity, and test use. Other ETS products are held to standards of fairness and quality for non-testing products.

I am proud of the ETS Standards, and I am committed to their continued implementation in our daily work and through ETS's audit program. I know that these Standards will help us greatly in our work to promote learning and performance, and in supporting education and professional development for all people worldwide.



Kurt M. Landgraf  
President and CEO

# Introduction

**Purpose.** The purposes of the ETS Standards are to help Educational Testing Service and subsidiary staff design, develop, and deliver technically sound, fair, and useful products and services, and to help auditors evaluate those products and services.

**Relation to Previous *ETS Standards for Quality and Fairness.*** This edition of the ETS Standards owes much to the earlier versions of the ETS Standards as first adopted by the Board of Trustees in 1981 and as updated in 1987 and 2000. The earlier Standards and the accompanying audit process stand as tangible evidence of ETS's dedication to quality and fairness, and of its willingness to be held accountable for meeting rigorous standards in its testing programs. The audit program established to monitor compliance with the original ETS Standards will continue to do so with the revised Standards.

This new version of the Standards maintains ETS's commitment to public accountability for quality and fairness, and extends that commitment to include customer service. In addition, the new Standards are applicable to all ETS products and services.

The 1987 edition of the ETS Standards had three levels of specificity: Principles, Policies, and Procedural Guidelines. (There were no statements specifically identified as "standards.") The new ETS Standards have two levels of specificity: (1) the essential goal of each chapter and (2) the standards themselves. The standards are followed by comments that elaborate on the standards and give guidance on how the standards are to be interpreted and applied.

**Application of the Standards.** The application of the ETS Standards will depend on the judgments of ETS staff and external evaluators. ETS intends the Standards to provide a context for professional judgment, NOT to replace that judgment.

No compilation of standards can foresee all possible circumstances and be universally applicable without interpretation. ETS does not intend the use of any of these standards to stifle adaptation to appropriate new environments, to slow the adoption of useful new assessment technologies, or to inhibit improvement.

If a consensus of sound professional judgments finds the literal application of a standard to be inappropriate in some particular circumstances, then the judgments should prevail. If staff cannot reach a consensus, they should request assistance from the Office of Corporate Quality Assurance to resolve the issue.

## Introduction

ETS does not always control all aspects of a product or service to which ETS staff contribute. ETS cannot force others who have *independent* control of an aspect of a product or service to comply with the ETS Standards. However, ETS should strongly encourage compliance. Whenever possible, adherence to the ETS Standards should be part of collaborative agreements.

**Relation to Joint Standards.** ETS strives to follow the *Standards for Educational and Psychological Testing*, issued jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Therefore, the ETS Standards are intended to be consistent with the relevant chapters of the Joint Standards, but have been tailored to ETS's specific needs and circumstances.

The ETS Standards focus more on desired outcomes and less on the specific means of reaching those outcomes than do the Joint Standards. The ETS Standards include material, such as that on customer service, not included in the Joint Standards. Furthermore, the ETS Standards exclude material found in the Joint Standards that is not relevant to ETS products or services.

**Future of the Standards.** ETS is committed to improving the Standards as the science of measurement advances, and as technology increases our capabilities. Please send suggestions for improving the ETS Standards to the Office of Corporate Quality Assurance.

# Overview

**Purpose.** The purpose of this overview is to provide a summary of the basic structure of the *ETS Standards for Quality and Fairness*.

**Structure.** There are 13 chapters following the introduction, this overview, and a discussion of the audit process. A glossary is provided to explain technical or specialized terms.

The following table summarizes the applicability of the standards in each chapter to assessment and non-assessment products and services.

## Applicability of Standards

Chapter	Non-Assessment Products & Services	Assessments
1 Developmental Procedures	✓✓✓	
2 Suitability for Use	✓✓✓	
3 Customer Service	✓✓✓	✓✓✓
4 Fairness	✓✓✓	✓✓✓
5 Uses and Protection of Information	✓✓✓	✓✓✓
6 Validity		✓✓✓
7 Assessment Development		✓✓✓
8 Reliability		✓✓✓
9 Cut Scores, Scaling, and Equating		✓✓✓
10 Assessment Administration		✓✓✓
11 Reporting Assessment Results		✓✓✓
12 Assessment Use		✓✓✓
13 Test Takers' Rights and Responsibilities		✓✓✓

The use of discrete chapters can be misleading in some respects. Fairness, for example, is a pervasive concern, and standards related to fairness could appropriately occur in every chapter. Placing the fairness standards in their own chapter was meant to stress the importance of the fairness standards, not to imply that they were isolated from other issues.

## Overview

The first two chapters of the ETS Standards apply to all products and services developed by ETS, except assessments. The next three chapters apply to all ETS products and services, including assessments. The next eight chapters apply primarily to assessments. The standards apply to all assessments, including paper-and-pencil assessments, computer-based assessments, portfolios, multi-media assessments, performance assessments, or other forms of evaluation developed by ETS. Some of the chapters (e.g., Reliability, Validity) are intended for readers with a knowledge of measurement.

Variations of some crucial standards are found in more than one chapter. For example, clearly stating the purpose of an assessment is central to many concerns, and related standards are included in several chapters. The documentation gathered to show compliance with one variation of the repeated standard will serve to show compliance with the other variations of the standard.

# Audit Requirements

The purpose of the audit requirements is to help ensure that ETS products and services will be evaluated with respect to a uniform, rigorous set of standards through a well-documented process.

Products and services developed by ETS must be periodically reviewed for compliance with the *ETS Standards for Quality and Fairness*. The ETS Standards are part of ETS's effort to ensure accountability for its products and services. In addition, a crucial part of the effort is the ETS Audit Program, through which the ETS Standards are applied systematically to ETS products and services.

The ETS Office of Corporate Quality Assurance establishes the audit schedules to ensure that ETS products and services are audited at reasonable intervals. In consultation with the Office of the President, the Office of Corporate Quality Assurance may extend the regularly scheduled audit cycle of once every three years. Postponements may be awarded based on excellent results in previous audits for products or services that are essentially unchanged since their last audit.

The ETS Office of Corporate Quality Assurance recruits auditors to perform each review. Auditors should reflect the diversity of ETS professional staff. The auditors assigned to a product or service must be trained in the audit process, be independent of the product or service being audited, and must, as a group, have the knowledge and experience necessary to make the required judgments about the product or service being evaluated.

Proper interpretation of ETS Standards depends on the judgments of professional staff. *The ETS Standards for Quality and Fairness are intended to guide auditors in the use of professional judgment, not to eliminate the need for it.*

The Office of Corporate Quality Assurance organizes audit teams to perform the reviews. Teams include individuals with sufficient technical knowledge and experience to assess products' and services' compliance with standards related to test development and statistical quality, and individuals with a wide range of general knowledge to assess areas of general accountability. Individuals from outside ETS also serve as members of some audit teams to provide fresh insights and public perspectives.

The Office of Corporate Quality Assurance trains auditors and program staff to perform their roles in the audit process using orientation sessions, written materials, reviews of previous audits, and mentoring.

## Audit Requirements

Program staff members evaluate the ways in which their products and services comply with each of the Standards. They assemble the documentation required to establish that the program's practices are reasonable in light of the Standards and present that documentation to the audit teams.

Auditors follow an audit model agreed upon by the program, the auditors, and the Office of Corporate Quality Assurance. Each audit team evaluates the product's or service's practices to determine whether or not they comply with the Standards. Whenever members of an audit team believe that a product or service does not comply with the Standards, they must explain why and make an appropriate recommendation for resolving the situation.

Programs should offer clients the opportunity to become involved in the audit process. Clients may choose to be involved in the audits in a number of ways, including participating in the program's self-audits and the preparation of audit materials, attending meetings with auditors, and discussing follow-up activities. In any case, regardless of the level of involvement they choose, clients are entitled to the results of the audit process and to information about the program's response to any requested changes.

The procedures used in reviews may vary. Auditors of a complex testing program may work in two teams, one addressing the less technical aspects of the program and the other addressing the more technical aspects. For other products and services, a single team of auditors may be preferable.

In some circumstances, the complexity of the testing program requires that program directors prepare large amounts of material in advance and mail it to each member of the audit team. In other cases, program staff and auditors review and discuss material in a single session.

In some audit models, program staff prepare written comments in advance of meeting with the auditors. In other models they do not. Audit models can differ by

- number of auditors;
- number of audit teams;
- amount of material sent to auditors in advance of the audit meeting;
- amount of elapsed time from start to finish of the audit; and
- number and length of meetings.

Participants in each audit, in consultation with staff in the Office of Corporate Quality Assurance, work together to select the most appropriate elements from the audit models to facilitate a thorough and efficient review.

Programs develop and implement action plans as necessary, possibly in collaboration with clients, to bring their product or service into full compliance with the ETS Standards. Cognizant officers, the Office of Corporate Quality Assurance, and the audit team evaluate the action plan.

If a program director and an audit team disagree on written comments or ratings, the program director may appeal the rating or comment to the Executive Director of the Office of Corporate Quality Assurance. Program directors may also request an exemption from follow-up action for good cause. However, few exemptions are granted. For an exemption to be granted, approval is needed from both the Office of the President and the program's cognizant officer.

The Office of Corporate Quality Assurance monitors program directors' progress in bringing their practices and policies into compliance with the Standards. Failure to correct deficiencies within a reasonable period may lead to discontinuance of the program.

The Office of Corporate Quality Assurance reports audit findings to the ETS Board of Trustees. Involvement of the Board of Trustees assures that the highest level of attention possible is paid to the results of the audits and to the quality of the entire process.

# 1

CHAPTER

## Developmental Procedures

### Purpose

**The purpose of this chapter is to help ensure that non-assessment<sup>1</sup> products and services will be developed and revised using planned, documented processes that include advice from diverse people, formative and summative evaluations, and attention to fairness and to meeting the needs of clients and users.**

This chapter applies to training, guidance, instructional, and other non-assessment products and services developed by ETS, intended for use outside of ETS. Products and services must be developed and maintained through procedures that ensure an appropriate level of quality, and attempt to satisfy client and user needs.

It is not the intent of this chapter to impose a single development procedure for all products and services. For example, some, but not all, new products are developed using the New Product Development Process.

### Standards

#### Standard 1.1

**Describe the intended purpose(s) of the product or service and its desired major characteristics.**

Provide sufficient detail to allow reviewers to evaluate the desired characteristics in light of the purpose(s) of the product or service.

#### Standard 1.2

**Describe the intended users of the product or service and the need that the product or service meets.**

The utility of products and services depends on how well the needs of the intended user groups are identified and met.

---

<sup>1</sup> See chapter 7 for analogous standards dealing with assessments.

## Chapter 1: Developmental Procedures

### **Standard 1.3**

**Document the development process used for existing products or services. Provide a plan for developing new products or services or revising existing ones.**

The documentation or plan for a product or service, including research projects, should address development procedures, schedules, staff with major responsibility for the project, internal and external reviews, and fairness and accessibility issues. For products and services developed for a particular client, work collaboratively with the client, as appropriate, to establish the development plan.

### **Standard 1.4**

**Obtain substantive advice and reviews from diverse internal and external sources, including clients and users.**

The development of a product or service, including research projects, must be informed by advice and reviews from diverse sources. As appropriate, include people representing different population groups, different institutions, different geographic areas, and so forth.

For products and services developed for a particular client, work collaboratively with the client to identify suitable reviewers. Obtain the reactions of current or potential clients and users, and the reactions of technical and subject-matter experts, as appropriate. Seek advice and reviews about fairness and accessibility issues and about legal issues that may affect the product or service.

### **Standard 1.5**

**Evaluate the product or service at reasonable intervals. Make revisions and improvements as necessary.**

Obtain judgmental and/or empirical data to demonstrate the appropriateness of the product or service and its utility to the intended users. Provide a schedule for the reviews and a rationale for the interval chosen. In general, five years should be the longest interval between evaluations.

# 2

CHAPTER

## Suitability for Use

### Purpose

**The purpose of this chapter is to help ensure that non-assessment<sup>2</sup> products and services are capable of meeting their intended purpose(s) for their intended population(s).**

This chapter generalizes the concept of validity to apply to training, guidance, instructional, and other non-assessment products and services intended for use outside of ETS.

The Standards do not require controlled experiments to prove that each product or service meets its intended purpose. There must, however, be documented logical and/or empirical evidence that the product or service should perform as intended for the appropriate population(s).

The amount and quality of the evidence required depends on the nature of the product or service and its intended use. As the possibility of negative consequences increases, the level and quality of the evidence required to show suitability for use should increase proportionately.

### Standards

#### Standard 2.1

**Describe how the product or service fits the ETS Mission and Values.**

Every product or service developed by ETS must further the ETS mission and must be congruent with ETS's values. The Office of Corporate Quality Assurance will provide the most recent editions of the Mission and Values Statements upon request.

#### Standard 2.2

**Obtain and document logical and/or empirical evidence that the product or service meets its intended purpose(s) for the intended population(s).**

Provide evidence that indicates that the product or service is likely to meet its intended purpose(s) for the intended population(s).

---

<sup>2</sup> See chapter 6 for analogous standards dealing with assessments.

## Chapter 2: Suitability for Use

The evidence could include such factors as the qualifications of the designers, builders, and reviewers of the product or service; the results of formative evaluations of aspects of the product or service in prototypes or pilot versions; and the opinions of subject-matter experts. Evidence based on controlled experiments is welcomed, but not required.

### **Standard 2.3**

**Establish and document procedures to maintain the technical quality, utility, and fairness of the product or service.**

Once a program establishes the technical quality, utility, and fairness of a product or service, the program must carry out procedures, such as periodic reviews, to ensure that the suitability is maintained. In general, five years should be the longest interval between reviews.

### **Standard 2.4**

**If relevant factors change, reassess the evidence that the product or service meets its intended purpose(s) for the intended population(s), and gather new evidence as necessary.**

Relevant factors include substantive changes in intended purpose, major changes to the product or service itself, and changes in the characteristics of the user population.

### **Standard 2.5**

**Provide information to interested users of the product or service to help them gather evidence that the product or service is meeting its intended purpose.**

Provide advice upon request concerning how to run local studies of the product's or service's effectiveness.

### **Standard 2.6**

**Warn intended users to avoid likely misuses of the product or service.**

No program can warn against (or even imagine) every misuse that might be made of a product or service. However, if there is evidence that misuses are likely (or are occurring), programs should warn users to avoid those misuses, and should inform users of appropriate uses of the product or service.

# 3

CHAPTER

## Customer Service

### Purpose

**The purpose of this chapter is to help ensure that meeting customer needs and maintaining high quality will be a major factor in designing, developing, and delivering products and services.**

The term “customer” includes clients, users of scores or other assessment results, purchasers of products and services developed by ETS, and test takers. Programs should identify their customers.

This chapter addresses the policies and practices that should be in place to ensure that ETS staff are aware of the needs of their customers and are striving to meet these needs.

### Standards

#### Standard 3.1

**Identify key clients and other customers, and obtain their input into product and service design, development, and operation.**

For products and services developed for a particular client, work collaboratively with the client as appropriate during the design, development, and operation of the products or services. On a regular basis, obtain information about the client’s desires and users’ needs, and how well each product or service meets them. For new products, this process should begin during the planning phase.

#### Standard 3.2

**Develop, in consultation with clients, service standards covering areas such as hours when phones will be answered, response times for inquiries, and speed of order fulfillment. Monitor the extent to which the standards are met.**

Where appropriate, review the service standards of other organizations providing the same or similar products and services.

### **Standard 3.3**

**Provide convenient methods for customers to obtain information about products and services, ask questions, make comments, or register problems or concerns. If an answer is required, respond promptly and courteously. Evaluate the effectiveness of the methods.**

Develop procedures to route inquiries to the appropriate staff for timely and accurate responses. Collect information on client and user satisfaction with interactions with staff. Develop action plans in collaboration with clients, as appropriate, to improve ineffective methods.

### **Standard 3.4**

**Periodically measure customer satisfaction. Use the information to attain higher levels of satisfaction.**

Obtain information from clients, institutions, and individuals concerning their satisfaction with products and services. For products and services developed for a particular client, work collaboratively with the client to do so.

The methods used for obtaining information can include formal surveys, focus groups, comment cards, customer advisory groups, complaint "hot lines," and so forth. Develop action plans in collaboration with clients, as appropriate, to remedy areas of low satisfaction.

### **Standard 3.5**

**Monitor the progress of work against schedules. Notify clients and others adversely affected if agreed-upon important deadlines will not be met.**

Try to minimize the problems for customers caused by the failure to meet important deadlines.

### **Standard 3.6**

**Verify that products or services conform to specifications or standards before they are released to customers.**

Independently recompute or inspect an appropriate sample of each product or service. Assess the reasonableness of results through review by technically competent staff. Review and proof printed materials. Test software.

**Standard 3.7**

**Verify and document the accuracy of internal products when the information is critical to external products.**

Early internal detection and correction of errors increases the likelihood that products reaching customers will adhere to specifications or standards.

**Standard 3.8**

**Correct any critical information found to be in error, if test takers or users of the product or service will be harmed by the error. Promptly distribute corrected information.**

Some critical errors may be discovered so long after their occurrence that correction is believed to serve no useful purpose. However, the permission of the cognizant ETS officer is needed to confirm that correction would serve no useful purpose.

## Fairness

### Purpose

**The purpose of this chapter is to help ensure that products and services will be designed, developed, and administered in ways that treat people equally and fairly regardless of differences in personal characteristics such as race, ethnicity, gender, or disability that are not relevant to the intended use of the product or service.**<sup>3</sup>

Products and services should take into account the diversity of the populations served. Generally, if sample sizes are sufficient, programs should investigate fairness at least for African American, Asian American, Hispanic American, and Native American (as compared to White) users of the product or service, and female (as compared to male) users of the product or service.

It is not feasible for programs to investigate fairness separately for all of the possible groups that may be defined. Therefore, programs should use experience or research to identify any other population groups to be included in evaluations for fairness. (In this chapter, the groups included in the evaluations of fairness are called the "studied" groups.)

There are many definitions of fairness in the professional literature, some of which contradict others. For the purposes of this chapter, fairness requires treating people with impartiality regardless of personal characteristics such as gender, race, ethnicity, or disability that are not relevant to their interaction with ETS. With respect to assessments, fairness requires that construct-irrelevant personal characteristics of test takers have no appreciable effect on test results or their interpretation.

ETS is responsible for the fairness of the products or services it develops and for providing evidence of their fairness. Users of a product or service are responsible for assessing the relevance of the evidence provided by ETS for their particular situations.

---

<sup>3</sup> For people with disabilities, equal treatment may involve providing accommodations appropriate to documented needs to give all people equal opportunities to use and benefit from the product or service.

# Standards

### Standard 4.1

**Address fairness in the design, development, administration, and use of the product or service, and document what was done. Provide a plan for addressing fairness for a product or service under development or facing major revision.**

Consult with clients as appropriate in the development of fairness plans. In either the documentation or the plan, as appropriate, demonstrate that reasonably anticipated potential areas of unfairness were or will be addressed. Some version of the fairness documentation or plan should be available for an external audience.

Group differences in performance do not necessarily indicate that a product or service is unfair. However, meaningful differences between groups in assessment results should be investigated to be sure they are not caused by construct-irrelevant factors.

Topics to be included in the documentation or the plan will vary depending on the nature of the product or service. *If it is relevant to the product or service (including assessments), and if it is feasible to obtain the data*, include information about the

- selection of groups to be studied;
- selection of variables to be studied;
- reviews done to ensure fairness;
- reactions of test takers or users of the product or service;
- appropriateness of materials for people with various backgrounds and characteristics;
- accessibility of the product or service;
- affordability of the product or service;
- evaluation of the linguistic or reading demands to ensure they are no greater than required;
- accommodations for people with disabilities; and
- effects of mode of presentation and response format.

In addition, for assessments, *if it is relevant for the assessment and feasible to obtain the data*, include information about the

- performance of studied groups, including evidence that different constructs may be measured in different populations;
- possible limitations to, or unintended consequences of, assessment use for studied groups;
- evidence of differential impact, differences in prediction as reflected in regression equations, or differences in validity evidence for studied groups;
- empirical procedures used to evaluate fairness (e.g., differential item functioning);

- comparability of different modes of assessments for studied groups;
- effects of time limits on studied groups;
- use of testing time, use of efficient test-taking strategies, or availability of coaching;
- different levels of experience with computers;
- choice of tasks within an assessment by studied groups;
- training designed to eliminate possible rater or administrator biases;
- criteria for computer scoring of complex responses;
- instructions given to test takers and raters;
- content balance of the assessment;
- accessibility of administration sites; and
- proper use and interpretation of the results for the studied population groups.

### **Standard 4.2**

**Obtain and document judgmental and, when possible, empirical evaluations of fairness for studied groups. As appropriate, ensure that various groups are represented. Ensure that symbols, language, and content that are generally regarded as sexist, racist, or offensive are eliminated, except when necessary to meet the purpose of the assessment, product, or service.**

Review materials, including written products, software, Web pages, and videos, to ensure that they meet the fairness review guidelines. Document the qualifications of the judges as well as the evidence they provide.

For assessments, when sample sizes are sufficient and the information is relevant, obtain and use empirical data relating to fairness, such as differential item functioning. If sufficient data are unavailable for some studied groups, provide an implementation plan for obtaining the data, if feasible. Use judgmental reviews whether or not empirical data are available.

### **Standard 4.3**

**Provide impartial access to products and services. For assessments, provide impartial registration, administration, and reporting of assessment results.**

Treat every user of products and services with courtesy and respect, and without bias, regardless of the user's race, gender, ethnicity, or other characteristics not relevant to the product or service offered.

### **Standard 4.4**

**When a construct can be measured in different ways that are reasonably equally valid, reliable, practical, and affordable, consider available evidence of group differences in assessment results in determining how to measure the construct.**

This standard applies when programs are developing new assessments or adding measures of new constructs to existing measures.

### **Standard 4.5**

**Provide appropriate and reasonable accommodations for people with disabilities, in accordance with applicable laws, ETS policies, and client's policies.**

When modifications are requested, collect substantiating evidence of the need for modifications. Provide the necessary accommodations at no additional cost to people with disabilities. The accommodations should be designed to ensure, to the extent possible, that the assessment measures the intended construct rather than *irrelevant* variance resulting from a person's disabilities.

If assessment content is changed, the modifications should be based on knowledge of the effects of disabilities on performance as well as knowledge of good testing practices.

Take the following actions, *as appropriate for the product or service*.

- If feasible, and if sufficient sample sizes are available, pilot test to evaluate the use of the product or service for people with specific disabilities.
- If feasible for assessments, and if sufficient sample sizes are available, obtain validity evidence for people with specific disabilities.
- Describe the modifications that will be made for people with disabilities, and the rationale for any such modifications.
- If scores are flagged, and if it is legally acceptable to do so, issue statements in manuals or other materials regarding the interpretation of assessment results.
- If feasible, and if sufficient sample sizes are available, use empirical information to help determine the modifications to be made.
- Specify the qualifications necessary for modifying, administering, and interpreting the results of modified products and services.
- Document the qualifications of the person or group recommending the specific modifications to be made.

### **Standard 4.6**

**If there is evidence that the construct may not be comparable between standard and modified assessments, flag scores.**

Consult with the client and the ETS General Counsel's Office as needed to provide score users with guidance on the proper interpretation of any scores resulting from modified assessments or administrations.

### **Standard 4.7**

**Consider the needs of nonnative speakers of English in the development and use of products or services. For assessments, reduce threats to validity that may arise from language differences.**

Take the following actions, *as appropriate for the product or service*.

- State the suitability of the product or service for people with limited English proficiency.
- If a product or service is recommended for use with a linguistically diverse population, provide the information necessary for appropriate use with nonnative speakers of English.
- If a translation is made, describe the process and evaluate the outcome and its comparability to the original version.
- If linguistic modifications are recommended, describe the modifications in a document available to the public.
- If an assessment is available in more than one language, and the different versions measure the same construct, attempt to administer the assessment in the individual's preferred language, unless knowledge of one of the other languages is to be assessed.
- When sufficient relevant data are available, provide information on the validity and interpretation of assessment results for linguistically diverse groups.
- If an interpreter is used, the person should be fluent in the source and target languages, experienced in translating, and have basic knowledge of the relevant product or service.

### **Standard 4.8**

**For research studies, perform separate analyses for studied groups when the information is relevant and sample sizes are sufficient. Obtain informed consent for participation of human subjects, and avoid negative consequences of participation for members of all groups.**

To be fair to research subjects, follow procedures approved by the Committee on Prior Review of Research. If there is reason to believe results may vary for members of different studied groups, design the research to allow appropriate analyses.

# 5

CHAPTER

## Uses and Protection of Information

### Purpose

**The purpose of this chapter is to help ensure that ETS will safeguard critical information, protect confidential information, and provide information to the public that allows evaluation of products and services and promotes their proper use.**

As appropriate, programs should provide information that promotes public understanding of measurement and related educational issues.

### Standards

#### Standard 5.1

**Provide potential users of products or services with the information they need to determine whether or not the product or service is appropriate for them.**

Provide information about the purpose and nature of the product or service, its intended use, and the intended population(s). The information should be available when the product or service is released to the public.

In addition, for assessments, provide examples of questions or tasks, descriptions of administration and scoring procedures, and information about the meaning of scores or other assessment results. If the assessment is available in varied formats such as computer-based and paper-and-pencil formats, provide information about the formats and the relative advantages and disadvantages of each.

#### Standard 5.2

**Use clear and jargon-free language in communications designed for a general audience.**

Explain any technical terms in language the audience is likely to understand.

### **Standard 5.3**

**Avoid making unsupported claims about the benefits of using the product or service.**

Claims made for improvement of performance, for example, should be supported by appropriate documentation. Review marketing materials to confirm their accuracy. Marketing materials and product names should accurately represent the product or service.

### **Standard 5.4**

**Retain the information necessary to verify assessment results for a period consistent with the intended use of the assessment. Maintain the confidentiality of personal or institutional information, and inform individuals that information about them will be kept confidential unless permission is obtained or disclosure is required by law. Inform individuals how long information about them will be reported.**

Information to support scores or other assessment results, such as answer sheets, should be retained for a reasonable period in case there is a need to check the accuracy of the reported results. Safeguard the confidentiality of information in paper, electronic (e.g., e-mail, voice mail, fax, computer memory, and so forth), and other formats during its creation, storage, and transmission. Use procedures such as encryption and firewalls as appropriate to safeguard confidential and proprietary information. Inform organizations receiving data of the need to uphold the confidentiality of information about individuals and institutions.

### **Standard 5.5**

**Ensure that information necessary for the operation of a program is safeguarded and recoverable in the event of a disaster.**

Provide disaster recovery procedures for crucial data and for key work processes required for the operation of a program. Establish plans for recovery to minimize the effects of the disaster on service.

### **Standard 5.6**

**Follow review procedures for research that will assure the research information is of high quality. Publish or otherwise disseminate research results in ways that promote understanding and proper use of the information, unless a justifiable need to restrict dissemination is identified.**

Obtain reviews, as appropriate for the research effort, of the

- rationale for the research;
- soundness of the design;
- thoroughness of data collection;
- appropriateness of the analyses; and
- fairness of the report.

Report research results in ways that minimize the possibility of misinterpretation and misuse.

### **Standard 5.7**

**Give non-ETS researchers reasonable access to ETS-controlled, nonproprietary information, if the privacy of individuals and organizations, and ETS's contractual obligations, can be met.**

Whenever possible, grant access to data facilitating the reanalysis and critique of published research.

### **Standard 5.8**

**Protect ETS's and clients' intellectual property rights with respect to such proprietary products as items, software, marketing studies, procedural manuals, new product development plans, and the like.**

Assure that staff, consultants, and external collaborators follow appropriate procedures to maintain copyrights, trademarks, trade secrets, and other proprietary rights. Consult the ETS Proprietary Rights Office for information about the best means of protecting intellectual property rights.

Some proprietary information may be disclosed in certain circumstances with proper safeguards. Consult the Proprietary Rights Office for advice before disclosing any proprietary information.

# 6

CHAPTER

## Validity

### Purpose

**The purpose of this chapter is to help ensure that programs will gather and document appropriate evidence for assessments to support the intended inferences and actions based on the reported assessment results.**

Validity is one of the most important attributes of assessment quality. Programs must provide logical and/or empirical evidence to show that each assessment is capable of meeting its intended purpose(s).

Validity is a unified concept, yet many different types of evidence may contribute to the demonstration of an assessment's validity. Validity is not based solely on any single study or type of evidence.

Though all types of evidence may contribute to the validation of any assessment, the type of evidence on which most reliance is placed will vary with the purpose of the assessment. The level of evidence required may vary with the potential consequences of the decisions made on the basis of the assessment's results.

Responsibility for validity is shared by ETS, by its clients, and by the people who use the scores or other assessment results. ETS will provide evidence of validity at least to the extent required by the following standards.

Users are responsible for evaluating the validity of scores or other assessment results used for purposes other than those specifically stated by ETS.

### Standards

#### Standard 6.1

**Describe how the assessment fits the ETS Mission and Values statements.**

Every assessment developed by ETS must further the ETS mission and must be congruent with ETS's values. The Office of Corporate Quality Assurance will provide the most recent editions of the Mission and Values statements upon request.

### **Standard 6.2**

**Clearly describe the construct (knowledge, skills, or other characteristics) to be measured, the purpose(s) of each assessment, the intended interpretation(s) of the scores or other assessment results, and the intended test-taking population(s). Make the information available to the public upon request.**

Validation efforts focus on an interpretation of the assessment results of some population for some particular purpose. Therefore, the validation process begins with complete and clear descriptions of what is to be measured (the construct), the purpose of the assessment, the intended interpretations of the scores or other results, and the population for which the assessment is designed. For some assessments, links to a theoretical framework are part of the information required as the validation process begins.

Because many labels for constructs, as reflected in names for assessments, are not precise, augment the construct label as necessary by specifying the aspects of the construct to be measured and those to be intentionally excluded, if any. For example, the construct label "verbal ability" could reasonably be construed to include aspects of speaking, listening, reading, and writing, but a particular assessment may include only measures of reading comprehension under that construct label. The program should specify the aspects of verbal ability measured by the assessment, and those aspects that have been intentionally excluded.

### **Standard 6.3**

**Provide a rationale for the types and amounts of evidence collected to support the validity of the inferences to be made on the basis of the assessment. For new assessments, provide a validity plan indicating the types of evidence to be collected.**

There should be a rationally planned collection of evidence relevant to the intended purpose of the assessment. If specific outcomes of assessment use are stated or strongly implied, include evidence to support the expectation of the outcomes. If a major line of validity evidence, such as criterion-related evidence, is excluded, describe the reasons for its exclusion.

The levels and types of evidence required for any particular assessment will remain a matter of professional judgment. Base the judgments on such factors as the

- intended inferences and actions based on the assessment results;
- intended outcomes of using the assessment;
- harmful actions that may result from an incorrect inference;
- likelihood that any incorrect inferences will be corrected before any harm is done;

- amount of research available on similar assessments used for similar purposes, in similar situations;
- technical feasibility of collecting data to address a particular aspect of validity;
- availability of appropriate criteria for studies of the predictive aspect of validity; and
- availability of sufficient samples of test takers for empirical studies of their performance.

### **Standard 6.4**

**Obtain and document the logical and/or empirical evidence that the assessment will meet its intended purpose(s) and support the intended interpretation(s) of assessment results for the intended population(s).**

The evidence should, as a whole, be sufficient to indicate that the assessment is capable of (1) meeting its intended purpose(s) and (2) supporting the intended interpretation(s) of assessment results for the intended population(s). Programs should compile the evidence into a coherent and inclusive rationale, or "validity argument," supporting the appropriateness of the inferences to be made on the basis of the assessment results.

Programs should investigate any clearly credible alternative explanations of evidence that might disconfirm the validity of the assessment.

Provide sufficient information to allow people trained in the appropriate disciplines to evaluate and replicate the data collection procedures and data analyses that were performed.

*If it is relevant to the validation argument for the assessment, and if it is feasible to obtain the data, provide information in the validity argument concerning the*

- characteristics of samples of test takers on which analyses are based, and how well the samples represent the population(s) of interest;
- procedures and criteria used to determine assessment content;
- qualifications of subject-matter experts, job incumbents, item writers, reviewers, and other individuals involved in any aspect of assessment development or validation;
- procedures used in any data-gathering effort, the conditions under which data were collected, the results of the data gathering (including results for studied subgroups of the population), the precision of reported statistics, and, if any adjusted statistics were reported, the unadjusted statistics on which they were based;

## Chapter 6: Validity

- cognitive processes used by test takers or by judges involved in scoring, if those processes are part of the intended construct;
- rationale or empirical basis for any computerized interpretations of assessment results;
- training and monitoring of scorers, or the algorithms used by automated scoring mechanisms;
- changes in test performance following coaching, if results are claimed to be essentially unaffected by coaching;
- statistical relationships among parts of the assessment, and among reported scores or other assessment results, including subscores;
- rationale and evidence for any suggested interpretations of responses to single items, subsets of items, subscores, or profile scores;
- relationships among scores or other assessment results, subscores, and external variables, including the rationales for selecting the external variables, their properties, and the relationships among them;
- characteristics of the criteria in predictive studies, including the extent to which the criteria reflect the domain of interest;
- relationships of the scores or other assessment results with the criteria, including other variables likely to be used in multiple prediction analyses;
- evidence that the assessment results are useful for assigning people to alternative placements, if common criteria are available;
- information about levels of criterion performance associated with given levels of assessment performance, if the assessment is used to predict adequate/inadequate criterion performance;
- general opportunity test takers have had to learn the content and skills measured by graduation or promotion assessments;
- characteristics and relevance of any meta-analytic evidence used in the validity argument; and
- evidence that the program's claims about the direct and indirect benefits of assessment use are supported, including claims suggested by the title of the assessment.

### **Standard 6.5**

**Warn potential users to avoid likely uses of the assessment for which there is insufficient validity evidence.**

No program can warn against (or even imagine) all the unsupported uses or interpretations that might be made of assessment results. However, experience may show that certain unsupported uses of the assessment results are likely. For example, licensing tests have been misused for job selection. If there is evidence that unsupported uses are likely (or are occurring), programs should warn users to avoid unsupported uses, and should inform users of appropriate uses of the assessment.

### **Standard 6.6**

**If the use of an assessment results in unintended consequences for a group that is studied, review the validity evidence to determine whether or not the consequences arise from invalid sources of variance. If they do, revise the assessment to reduce, to the extent possible, the inappropriate sources of variance.**

Appropriately used scores or other assessment results on a valid assessment may have unintended consequences. For example, a valid licensing test may have differential impact for some groups. Unintended consequences do not necessarily invalidate the use of an assessment. It is necessary, however, to investigate whether the unintended consequences may be linked to non-construct-related factors or to construct underrepresentation. If so, take corrective actions.

### **Standard 6.7**

**If relevant factors change, reevaluate the evidence that the assessment meets its intended purpose(s) and supports the intended interpretation(s) of the assessment results for the intended population(s), and gather new evidence as necessary.**

Relevant factors include substantive changes in intended purpose, intended interpretation of assessment results, format, assessment characteristics, administration or scoring modes, population characteristics, or the domain assessed.

There is no set time limit within which programs must reassess the validity evidence. In general, however, five years should be the maximum period that goes by without a reassessment of the validity evidence.

### **Standard 6.8**

**Provide advice to users of scores or other assessment results to help them gather and interpret their own validity evidence.**

Tell users of assessment results how to conduct and interpret local validity studies. Users are responsible for validating the interpretations of assessment results if the assessments are used for purposes other than those explicitly stated by ETS.

# 7

CHAPTER

# Assessment Development

## Purpose

**The purpose of this chapter is to help ensure that assessments will be constructed using planned, documented processes that include advice from diverse people; formative and summative evaluations; and attention to fairness, reliability, and validity.**

Developers should work from detailed specifications, obtain reviews of their work, use empirical information when it can be obtained, and evaluate their finished products.

The standards do not require that the same developmental steps be followed for all assessments.

## Standards

### Standard 7.1

**Obtain or develop documentation concerning the intended purpose(s) of the assessment, the population(s) to be served, and the construct(s) to be measured.**

Developers must know what the assessment is intended to measure, the characteristics of the intended test takers, and how the information derived from the assessment is intended to be used. For some programs, the information has been collected and need not be re-created. For other programs, obtaining the information may be part of the developers' task. If the information has to be obtained, work collaboratively with clients as appropriate.

### Standard 7.2

**Document the desired attributes of the assessment in detailed specifications. Document the rationales for major decisions, and document the process used to develop the specifications.**

Assessment developers need detailed blueprints for constructing assessments. Include, *as appropriate for the assessment*, information about the

- content and skills or attributes to be measured, including detailed descriptions, and, where relevant, critical content to be included and content to be excluded, or tasks to be performed;
- outcomes of job or curriculum analyses;
- intended test-taker population(s), including major population groups;
- uses of assessment results and whether the interpretation of those results is to be relative or absolute;
- administration mode;
- traditional or automated item-selection rules;
- need to cross-validate empirical item selections;
- rules for including material representing various population groups;
- item types and numbers of each item type;
- rules for sequencing items;
- relative weights to be given to each part of the domain that is to be measured (including item weights, if any);
- timing, and whether or not the assessment is intended to be speeded;
- directions for test takers;
- intended level of difficulty, and the target distribution of item difficulties;
- mean discrimination index;
- requirements regarding the item response model, calibration procedures, and item parameters;
- target information curves;
- procedures for branching decisions, for terminating the administration, and for scoring adaptive assessments;
- scoring methods and rubrics, and guidelines for selecting, training, monitoring, and evaluating scorers;
- automated scoring or score interpretation algorithms;
- requirements for modified assessments for people with disabilities;

- requirements regarding equating, including the content and statistical specifications for equating items; and
- rules for using differential item functioning or other group-related data.

### **Standard 7.3**

**Write items that meet generally accepted guidelines. Assemble assessments that meet specifications or write automated assembly rules that result in assessments that meet specifications. Document the development process.**

Appropriate guidelines for writing items and assembling assessments are included in the ETS Test Creation policies and procedures documents. Items should be linked to specifications and should not introduce construct-irrelevant variance. Assessment content should represent the defined construct and exclude sources of variance unrelated to the construct. It should ensure, to the extent possible, that the intended interpretations of test results are supported.

### **Standard 7.4**

**Obtain internal and/or external reviews of the assessment and related materials. Document the qualifications of the reviewers and the results of their reviews.**

Obtain, *as appropriate for the assessment*, reviews of the

- specifications and their links to the intended interpretations of assessment results;
- items, including appropriateness for studied population groups;
- links of items to specifications and/or to occupational tasks;
- assembled assessments, or assembly rules and samples of assembled assessments;
- directions for test takers;
- ancillary materials such as descriptive booklets and test-preparation materials;
- scoring rubrics and procedures, and training materials for scorers;
- automated scoring or score interpretation rules; and
- extent to which the assessment represents the defined construct and excludes sources of variance unrelated to the construct.

## Chapter 7: Assessment Development

Obtain substantive contributions from qualified persons who represent relevant perspectives, professional specialties, and population groups, and users of the results of the assessment. For assessments developed for a particular client, work collaboratively with the client to identify appropriate reviewers. Include, as appropriate, reviewers who are not members of the ETS staff.

The reviews of items, directions, and ancillary materials must be performed by people who are familiar with the specifications and purpose of the assessment, and with the characteristics of its intended population.

Important aspects of the review include content accuracy, suitability of language, match of items or tasks to specifications, fairness for population groups, editorial considerations, completeness and clarity of directions and sample items, adequacy of scoring rubrics, appropriateness of presentation and response formats, and appropriateness of difficulty. For test-preparation materials, an important aspect of the review is ensuring that use of the materials will not impair the validity of the assessment.

For some subject areas, separate reviewers will be needed for content accuracy and for the other aspects of item quality.

### **Standard 7.5**

**Pretest items if feasible. If pretesting is not feasible, use collateral information about the items, review the results of administering similar items to similar populations, and/or conduct a preliminary item analysis before scores or other assessment results are reported.**

When sample sizes are sufficient to permit meaningful analyses, and there is reason to believe the information will be useful, obtain data on item performance of relevant population(s). Document the sample-selection process and the characteristics of the resulting sample(s).

Not every program can pretest items in advance of operational administration. Programs that are unable to pretest should try to use collateral information (e.g., number of operations and level of cognition required to respond, closeness of the distracters to the key, performance of similar items used with similar populations) to help predict the items' characteristics.

### **Standard 7.6**

#### **Evaluate the performance of assessments after operational administration.**

Carry out timely and appropriate analyses, including analyses for reliability, intercorrelation of sections or parts, and indications of speededness. Adaptive assessments may carry out evaluations of sample forms after simulated administrations, but data on real test takers should be evaluated when possible.

Evaluate representative test editions in terms of the degree to which they meet their psychometric specifications. If feasible and appropriate, investigate the sensitivity of performance to practice and coaching. When sample sizes are sufficient to permit meaningful analyses, and there is reason to believe the information will be useful, obtain data on the performance of studied population groups.

### **Standard 7.7**

#### **Periodically review active items, assessments, and ancillary materials to ensure that they continue to be appropriate and in compliance with current applicable guidelines.**

Programs should determine, in collaboration with clients as appropriate, review periods depending on the nature of the assessment. Generally, however, five years should be the longest interval between reviews.

For computer-adaptive tests, periodically evaluate, as appropriate, the adequacy of the fit of item response models, the size and security of item pools, and the adequacy of the algorithms used to select items.

Evaluate the continuing accuracy, adequacy, and fairness of the content specifications, items, directions, descriptive materials, practice materials, and human or automated scoring procedures.

### **Standard 7.8**

#### **Protect the security of confidential assessment materials throughout the development process.**

Follow documented procedures for storing, using, and transporting confidential materials in traditional and in electronic media.

# Reliability

## Purpose

**The purpose of this chapter is to help ensure that scores or other reported assessment results will be sufficiently reliable to meet their intended purposes, and that programs will use appropriate procedures for determining and reporting reliability.**

Reliability refers to the extent to which scores (or other reported results) obtained on a specific form of an assessment, administered at some particular time, and possibly scored by some particular rater(s) can be generalized to scores obtained on other forms of the assessment, administered at other times, and possibly scored by some other rater(s).

Reliability can also be viewed as an indicator of the extent to which assessment results are free from the effects of random variation caused by such factors as the form of an assessment, the time of administration, or the scorers.

It is not the purpose of this chapter to establish minimum levels of reliability, nor to mandate the methods by which programs estimate reliability for any particular assessment.

## Standards

### Standard 8.1

**Ensure that any reported scores (including subscores), or other assessment results, are sufficiently reliable to support their intended interpretation(s).**

The level of reliability required for an assessment is a matter of professional judgment. If a wrong decision based on a reported score or other assessment result has negligible consequences, low levels of reliability may be acceptable.

### Standard 8.2

**Estimate reliability using methods that are appropriate for the characteristics of the assessment and the intended use(s) of the results. Use methods that take into account important sources of possible variation in assessment results.**

Different types of assessments require different methods of estimating reliability. *If it is relevant for the type of assessment or type of reported assessment results, and if it is feasible to obtain the data,*

- take account of the existence of the different factors in calculating reliability for multifactorial assessments;
- use test-retest or alternate form reliability estimates for speeded assessments;
- calculate both inter-rater reliability and reliability of test-taker performance for subjectively scored assessments;
- provide alternate form reliability estimates for adaptive assessments;
- report standard errors for the level of aggregation at which assessment results are reported; and
- take the sampling scheme into account when estimating reliability for assessments using matrix sampling.

Different ways of estimating reliability may include different sources of variation (e.g., form-to-form differences, scorer differences, or differences in the performance of the test taker over time). Consistency in one source of variation (such as agreement among scorers of the same task) does not imply consistency in other sources of variation (such as test-taker consistency from task to task).

An assessment's scores or other results could have many reliabilities, depending on factors such as the method used to estimate reliability, the sample of test takers, or the part of the score scale at which decisions are being made. Do not assume that estimates of reliability derived using different procedures are necessarily equivalent.

### Standard 8.3

**Provide information that will allow users of assessment results to judge whether reported assessment results (including subscores) are sufficiently reliable to support their intended interpretation(s). When appropriate, include estimates of errors of measurement for the reported assessment results. If users are to make decisions based on the differences among scores, subscores, or other assessment results, provide information on the reliabilities and standard errors of those differences.**

Provide score users with information that will enable them to evaluate the reliability of the assessment results they are using. Include, *as appropriate for the assessment*, reliability estimates, information functions, standard errors of measurement, and conditional standard errors of measurement. Technical publications should include standard errors in raw score units as well as in scaled score units.

Inform score users about the sources of variation in scores or other assessment results considered significant for interpretation of the results, such as form-to-form differences in content or the degree of subjectivity in the scoring process.

If it is feasible to obtain the data, provide information about the reliability or consistency of the pass-fail decisions of tests used with cut scores. When appropriate, provide information about the standard error of measurement or other similar coefficients around the cut score.

### Standard 8.4

**Document the reliability analyses. Provide sufficient information to allow knowledgeable people to evaluate the results and replicate the analyses.**

*If it is relevant to the reliability analyses performed for the assessment, and if it is feasible to obtain the data, provide information concerning*

- the method(s) used to assess the reliability of the scores or other assessment results, the rationale for using them, the formula(s) used, and/or appropriate references;
- the conditions under which the data were collected;
- the population involved (for example, demographic information, education level, employment status);
- the selection procedures used and the characteristics of any samples for which reliability is estimated, including summary score statistics;

## Chapter 8: Reliability

- the major sources of variation in scores or other assessment results accounted for and not accounted for in the reliability analysis;
- the variance associated with each of the major sources of variation in scores or other assessment results accounted for by the reliability analysis;
- a reliability coefficient, an overall standard error of measurement, an index of classification consistency, an information function, or other information about the consistency of scores or other assessment results;
- conditional standard errors of measurement or other measures of consistency for score regions within which decisions about individuals are made;
- the time interval between administrations, and the order in which the forms were administered, if an alternate forms method was used;
- indications of the speededness of assessments;
- the effects of response modes that may be unfamiliar to test takers;
- reliabilities for students in different grades, if the assessment is used in a wide range of grades;
- any procedures used for judgmental or automated scoring;
- the level of agreement among independent judgmental scorings, by location if scoring is done at multiple sites; and
- the adjusted and unadjusted coefficients, if reliability estimates are adjusted for restriction of range.

### Standard 8.5

**If it is feasible to obtain adequate data, conduct separate reliability analyses whenever significant modifications are permitted in the assessment or conditions of administration or scoring.**

If assessments are administered in long and short versions, report reliability for both versions. To accommodate test takers with disabilities, programs may extend time limits or provide modified editions of the assessment. If feasible, aggregate data to allow the eventual assessment of reliability and/or standard errors of measurement for the different types of nonstandard administrations.

**Standard 8.6**

**Evaluate the reliability and standard error of measurement of reported assessment results for studied population groups, if the need for such studies is indicated and if it is feasible to obtain adequate data.**

Reliability estimates may differ for different groups. If sample sizes are sufficient, and if there is reason to believe the information will be useful in light of the intended use of the scores or other assessment results, investigate the reliabilities of assessments for studied population groups. Take differences in group variability into account in evaluating differences in group reliability.

# Cut Scores, Scaling, and Equating

## Purpose

**The purpose of this chapter is to help ensure that assessments will use score-reporting scales that are meaningful. If ETS participates in a cut-score study, it will use rational, clearly described procedures. Assessments that are meant to be linked to other assessments will have a level of comparability commensurate with the use(s) made of the scores.**

It is not the purpose of this chapter to specify the scales that programs may use, nor to require any particular method of equating, nor to require any particular method of setting cut scores.

## Standards

### Standard 9.1

**Use reporting scales for scores or other assessment results that are appropriate for the intended purpose of the assessment and that discourage misinterpretations. Provide a rationale for the reporting scale(s) selected.**

For established programs, the rationale may simply be the importance of maintaining the continuity of the historical scale.

### Standard 9.2

**If results on different assessments or on different forms of an assessment are to be treated as though they were equivalent or comparable, use appropriate linking or equating methodologies.**

There are different ways to link scores or other assessment results, ranging in rigor from strict equating models to judgmental methods. Any of the levels of linking may be acceptable in appropriate circumstances. Define the equating or other linking procedures used, and indicate why the program selected the procedure used.

### Standard 9.3

**Describe the equating or other linking studies that were done in sufficient detail to allow knowledgeable people to evaluate and replicate the studies. Describe the limitations of linking studies done when the linked assessments are not alternate forms of the same assessment.**

*If relevant to the procedures used, and if it is feasible to obtain the data, provide information about such factors as the*

- assumptions underlying the methods used;
- characteristics of the test-taking population;
- characteristics of the sample(s) selected and relevant sample statistics;
- reasons for considering groups of test takers to be equivalent, if the equating study relies on the equivalence of the groups;
- content and statistical characteristics of linking items or anchor tests, if the equating study makes use of them;
- similarity of linking items or anchor tests to the forms being equated;
- statistics generated in the study, including standard errors of equating or their equivalents, when feasible;
- similarity of content and the statistical relationships among the forms being linked; and
- adequacy of the fit of the model to the data when methods of equating based on item response theory are used.

### Standard 9.4

**Check the stability of the reporting scale whenever assessment or population characteristics change significantly. If scores or other assessment results are no longer equivalent, take steps to minimize misinterpretations. Check the stability of the reporting scale periodically, if appropriate use of the scores or other assessment results depends on the stability of the scale.**

If a change to the assessment or to the population changes the meaning of the assessment results, one alternative would be to change the scale to minimize confusion between the old and new meanings of the results. Another alternative would be to communicate the differences clearly to score recipients.

Select an appropriate schedule for checking reporting scales for stability, and provide a rationale for the interval chosen.

### **Standard 9.5**

**If ETS is involved in designing a cut-score study, provide users with the information they need to choose an appropriate methodology, and describe the logical and/or empirical evidence supporting the classifications made on the basis of the cut score.**

Generally, cut scores must be set by authorities who have the responsibility to do so under the laws of some jurisdiction. Therefore, ETS does not set cut scores. ETS may, however, design and implement standard-setting studies at a client's request, and provide other data, such as score distributions, when available to assist the client in setting a cut score.

The "validity argument" described in standard 6.4, combined with a description of the cut-score study, should serve as appropriate documentation to support the classifications made on the basis of the cut score(s).

### **Standard 9.6**

**For a cut-score study, use an appropriate data-gathering plan, choose appropriate sample(s) of raters from relevant populations, and train the raters in the method they will use. Document the study in sufficient detail to allow knowledgeable people to evaluate and replicate the study.**

When feasible, set cut scores using empirical information concerning the relation between assessment performance and relevant criteria. (In many circumstances, such information may be unobtainable.)

Ensure that the raters in a cut-score study understand the purpose of the assessment and how to apply the cut-score process that is to be used. The raters should have a sound basis for making the required judgments. If the data are available, provide estimates of the effects of setting the standard at various points.

The documentation of the cut-score study should include information about how and why the raters were selected, and how they were trained. Document how the individual judgments were combined. Include full descriptions of the procedures followed and the results. When feasible, provide estimates of the variation that might be expected in the cut score.

# 10

CHAPTER

## Assessment Administration

### Purpose

**The purpose of this chapter is to help ensure that assessments will be administered in an appropriate manner that provides accurate, comparable, and fair measurement for each test taker.**

Administration procedures, including the level of security required, may vary with the type and purpose of the assessment. For assessments developed for a particular client, collaborate with the client, as appropriate, to ensure proper assessment administration.

This chapter is not meant to specify the exact procedures for any assessment administration.

### Standards

#### Standard 10.1

**Provide those who administer assessments with timely, clear, and appropriate information about administration procedures.**

Develop, in collaboration with clients as appropriate, clear, written administration procedures. Give the people who administer assessments information about the

- purpose of the assessment, the population to be tested, and the information needed to respond to expected questions from test takers;
- qualifications required to administer the assessments;
- required identifying information for test takers, admittance procedures, timing, and directions;
- materials, aids, or tools that are required, optional, or prohibited;
- maintenance of appropriate security procedures;
- actions to take if irregularities are observed;
- operation of equipment and software as needed for the administration;

## Chapter 10: Assessment Administration

- provision of approved accommodations to test takers with disabilities; and
- resolution of problems that may delay or disrupt testing.

Present this information to people who administer assessments through training, instruction manuals, videotapes, periodic updates, or other materials in a reasonable time before the administration date.

### **Standard 10.2**

**Tell test takers what they can expect in the registration and reservation processes, and at the administration.**

Provide clear information to test takers, including how to

- register for the assessment;
- request accommodations for people with disabilities;
- present appropriate identification materials;
- distinguish among the materials, tools, and aids that are required, optional, or prohibited during the assessment;
- take the assessment and make responses;
- observe the rules dealing with proper behavior during the administration; and
- report problems or complain about registration and reservation services, the administration, and/or the items.

Warn test takers of the consequences of misconduct.

### **Standard 10.3**

**Provide a reasonably comfortable testing environment in locations reasonably accessible to the intended population(s). Monitor administrations as appropriate to ensure relevant procedures are followed.**

Administer assessments in an environment with appropriate temperature control, reasonable furniture, adequate lighting and work space, low noise levels, and few disruptions. As necessary, make accommodations to provide access to a testing site for people with disabilities. Monitor samples of administration sites, as feasible, to ensure that assessments are being administered as specified.

### **Standard 10.4**

#### **Protect secure assessments.**

The level of security needed to protect the integrity of the assessment and scores or other assessment results depends on the purpose and nature of the assessment. Security is not a concern for some assessments.

For secure assessments, whether administered by computer or printed booklet, maintain security in all phases of the administration process, including distributing materials to administration sites, operating administration sites, storing materials, and returning materials.

### **Standard 10.5**

#### **Eliminate opportunities for test takers to attain scores or other assessment results by fraudulent means, to the extent possible.**

For secure assessments, discourage impersonation attempts by requiring test-taker identification procedures. To discourage copying, follow program seating guidelines for arranging seating at administration sites. People who administer assessments should monitor test takers and take appropriate action if irregularities are observed. Work with clients to establish appropriate security procedures.

# 11

CHAPTER

## Reporting Assessment Results

### Purpose

**The purpose of this chapter is to help ensure that ETS will provide correct, understandable scores or other assessment results and appropriate interpretive information to recipients.**

Scores or other assessment results must be accurate and communicate meaningfully with the intended recipients.

It is not the purpose of this chapter to limit the specific ways in which assessment results for individuals or groups should be reported.

### Standards

#### Standard 11.1

**Establish, document, and follow procedures to ensure the accuracy of scoring. Monitor scoring errors and correct sources of error.**

As appropriate, check samples of scores or other assessment results for correctness. For example, hand score scanned answer sheets, review response patterns leading to computer-generated scores, compare computer-generated scores of constructed-response items with those generated by human scorers, and have independent re-scores of scores based on human judgment. Human scorers should have specified criteria for scoring and be trained in their application.

### **Standard 11.2**

**Explain the meaning of the score scale to recipients. Provide users with the information they need to evaluate the characteristics and meaning of the scaled scores, or any reported raw scores, or other assessment results.**

Possible recipients of assessment results such as scores, score profiles, aggregated data, and imputed scores include test takers, parents, teachers and other school personnel, admissions and placement officers, colleges, employers, and agencies. Different recipients may require different types of information at different levels of technical detail.

For some programs, summaries of scores or other assessment results are released to the media. For programs in which there is a public interest in the data, provide information that will help the news media and the general public understand the scale on which the results are reported.

### **Standard 11.3**

**Provide information that minimizes the possibility of misinterpretation of individual assessment results, or results for groups. Warn intended recipients of limitations and likely misinterpretations of the reporting scale.**

If raw scores or percent correct scores are reported, inform intended recipients of the limitations of such scores with respect to comparability across assessments. If scores are reported as labels, use the least stigmatizing labels consistent with accuracy. If pass-fail scores are reported, give failing test takers some information about their performance relative to the cut score, when it is feasible and helpful to test takers.

When the misinterpretations of score scales are likely, warn score users to avoid those misinterpretations. For example, scores for different assessments may be reported on the same scales, although the program may not have done anything to ensure comparability of the scores. Caution users that such scores are not interchangeable.

### **Standard 11.4**

**Provide recipients with an appropriate frame of reference for evaluating the performance represented by test takers' scores or other assessment results.**

The frame of reference might include information based on norms studies, carefully selected and defined program statistics, research studies, or logical analysis.

Clearly describe the nature of the group(s) on which the information is based. Distinguish among carefully sampled representative norm groups and groups of people who choose to take the assessment.

### **Standard 11.5**

**If the program reports normative scores, select norm groups that are meaningful for score users. Describe the norm group(s) and the norming studies in sufficient detail to allow score users to determine the relevance of the norms for local test takers, and to allow knowledgeable people to evaluate and replicate the studies. Update the norming studies if the results become outdated or misleading.**

Descriptions of norming studies should include the dates of the studies, definition of the population(s) sampled, the procedure used to draw the samples, participation rates, and any weights used to make the sample data better represent the population. Include an estimate of the precision of the norms.

Program norms based on the people who choose to take an assessment can be helpful, but inform score users of the limitations of such norms.

If published norms are likely to be inappropriate for local test takers, warn score users to evaluate the relevance of the published norms. If local norms are necessary to support the intended use of the scores, either provide the local norms or tell recipients how to develop local norms.

Set a schedule for the review of normative data and provide a rationale for the selected review interval.

### **Standard 11.6**

**Present score information or other assessment results about population subgroups in a way that encourages correct interpretation and use.**

Sufficient data should be available for each of the subgroups under consideration to make the information useful. Avoid developing separate information for groups so small as to make the data potentially misleading. Accompany the information with a carefully described rationale for interpreting and using it.

# 12

CHAPTER

## Assessment Use

### Purpose

**The purpose of this chapter is to help ensure that ETS will provide information that describes and encourages proper assessment use. ETS will warn intended users of assessment results to avoid common misuses of the assessment.**

ETS will promote proper use of assessments and help score recipients use assessments fairly and appropriately, in accordance with supporting evidence.

### Standards

#### Standard 12.1

**Provide intended users of assessments with the information they need to evaluate the appropriateness of assessments. Provide users with opportunities to consult with ETS staff about appropriate uses of assessments.**

Provide users of assessments with information such as the

- intended purpose(s) and population(s) of the assessments;
- content and format of the assessments;
- difficulty, reliability, and validity of the assessments;
- availability and applicability of normative data;
- administration and scoring requirements;
- policies for data retention and release; and
- representative relevant research.

The information should be at a level that is likely to be understood by the intended users. Tell users how to reach appropriate ETS staff for answers to questions about uses of assessments.

### **Standard 12.2**

**Encourage proper assessment use and appropriate interpretations of assessment results. Caution users to avoid common, reasonably anticipated misuses of the assessment.**

Encourage proper assessment use by taking such actions as informing assessment users

- how to use scores or other assessment results in conjunction with other relevant information (if such information is available and useful);
- how to evaluate score differences between individuals (or between subscores for the same individual);
- of alternative plausible explanations for poor performance;
- of the need to allow students who must pass a test to be promoted or granted a diploma, or individuals seeking certification, a reasonable number of opportunities to succeed; and
- of the need to become familiar with the responsibilities of assessment users as described in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999).

Discourage misuses of scores or other assessment results by warning users of likely problems. Tell users what to do to avoid the problems.

### **Standard 12.3**

**Investigate credible allegations of assessment misuse, when feasible. If assessment misuse is found, inform the client and the user. Inform the user of the appropriate use of the assessment. If the misuse continues, consult with the client concerning appropriate corrective actions.**

The appropriate actions may include withholding scores or other assessment results from score recipients who persist in harmful misuse.

### **Standard 12.4**

**Provide information and advice to help interested parties evaluate the appropriateness, utility, and consequences of the decisions made on the basis of assessment scores or other assessment results.**

Score users, policymakers, and clients are among the interested parties to consider. Relevant and credible information about the likely effects of assessment use may be of great help to policymakers who are considering mandating assessments for some purpose.

### **Standard 12.5**

**Give people outside of ETS who control aspects of administration and scoring the information they need to perform those responsibilities correctly.**

Administration, scoring, and reporting of ETS-developed assessments are often controlled by ETS. In cases where they are not, ETS should tell those who have control of one or more aspects of administration, scoring, or reporting about appropriate policies and actions to protect test takers, institutions, and the public. If relevant, provide information about maintaining the security of the assessment, administering the assessment, releasing useful summaries of scores or other assessment results, protecting test-takers' and institutions' rights to privacy, and allowing test takers to challenge the correctness of scores.

### **Standard 12.6**

**Establish a policy to deter use of obsolete scores or other assessment results. State the time during which assessment results will continue to be reported.**

Some scores or other assessment results are likely to become obsolete rather quickly. For example, competence in English as a second language may change rapidly with immersion in an English-speaking culture. Other scores are likely to be meaningful for longer periods. All scores, however, are likely to become obsolete after some amount of time. Therefore, report the date of the administration with the score or other assessment results, unless there is a good reason not to do so.

# 13

CHAPTER

## Test Takers' Rights and Responsibilities

### Purpose

**The purpose of this chapter is to help ensure that ETS will make test takers aware of their rights and responsibilities, and will protect their rights during all phases of the assessment process.**

Many of the rights of test takers are detailed in other chapters, such as Fairness, Assessment Administration, and Reporting Assessment Results, and are not repeated in this chapter.

### Standards

#### Standard 13.1

**Inform test takers of their rights and responsibilities.**

Statements of the rights of test takers should include such issues as

- their right to information about the nature and purpose of the assessment;
- any recourse they have if the assessment or the administration is flawed; and
- whether or not scores may be canceled by the test taker or by ETS, and if so, under what conditions.

Generally, test takers and/or their legal representatives are entitled to a copy of the test scores or other assessment results reported by ETS, unless they have been canceled or the right has been waived.

Statements of the responsibilities of test takers should include such issues as

- preparing for the test appropriately;
- following program requirements;
- not copying or taking secure materials;
- not copying responses from others, not using unauthorized materials, not representing someone else's work as their own; and
- not knowingly facilitating copying of responses or taking of secure materials by others.

### Standard 13.2

**Provide all test takers with respectful and impartial treatment, appropriate access to assessment services, and information about the assessment and the administration process.**

Answer reasonable questions from test takers before the administration. Tell test takers in advance if they will not be permitted to ask questions during any subsequent phase of the testing process.

Inform test takers (or their parents or guardians) of such issues as

- the characteristics of the assessment, including any novel item or administration formats;
- the testing fees, if any;
- the intended uses of any scores or other assessment results, as well as the conditions under which results will be reported and to whom;
- the score or combinations of scores required to pass assessments used with passing scores, if the information is available;
- the advantages and disadvantages of each mode of assessment, if multiple modes are offered;
- which materials are required for the assessment, which are optional, and which are prohibited;
- whether the scores will be flagged, if a test taker has a modified administration;
- appropriate test-taking strategies;
- how often, how soon, and under what conditions the assessment may be taken again;
- whether there are scoring procedures that may affect their results (e.g., if there is a correction for guessing), unless providing such information is inconsistent with the purpose for testing;
- score cancellation policies in the case of irregularities;
- opportunities for test takers to cancel scores;
- whether test takers will have access to copies of editions of the assessment or samples of items, and whether it is possible to obtain a record of responses or have the assessment rescored; and
- if specialized equipment is used in the testing, how to obtain practice in its use before testing, unless evaluating the ability to use the equipment is intended.

### **Standard 13.3**

#### **Obtain informed consent from test takers as necessary.**

Administer assessments or release personally identifiable scores or other assessment results only when test takers (or their parents or guardians) have provided informed consent, except in circumstances in which consent is clearly implied, or when testing without consent has been mandated by law or government regulation. Other exceptions to the need to obtain consent occur when testing is conducted as a regular part of school activities, or when confidentiality of data is ensured. Inform test takers if personally identifiable information about them will be used for research purposes.

### **Standard 13.4**

#### **Tell test takers how to register complaints about items believed to be flawed, assessment content believed to be inappropriate, administration procedures believed to be faulty, or scores or other assessment results believed to be incorrect.**

Respond in a timely fashion to complaints. Some identified problems may require the reporting of revised scores or other assessment results, or the offer of a re-assessment.

### **Standard 13.5**

#### **If ETS cancels scores or does not report scores within the normally expected reporting time, follow documented procedures designed to protect the rights and privacy of test takers whose scores are questioned on the grounds of irregularity or misconduct. Allow test takers to offer relevant evidence when appropriate, and consider such evidence in deliberations about the questioned scores.**

Programs with secure assessments should have documented procedures, approved by the client and the General Counsel's Office, describing procedures in the case of possibly invalid scores. If a test taker's scores are under review and may be canceled or withheld, perform the review in a timely manner.

When appropriate, inform test takers of the procedures that will be followed, the standard of evidence that will be used to determine whether to cancel or withhold scores of suspect validity, and the options that test takers may select for resolving the matter. When applicable, consider reasonably available relevant information that includes material the test taker may choose to provide.

# Glossary

**Absolute Score** - A score that is interpreted without reference to a distribution of scores. An absolute score is interpreted as being either above or below some standard. For example, a score on an assessment with a passing score set at 80 percent of the questions correct, without basing the decision on how many people will score above or below that point, is an absolute score. See **Cut Score, Standard**.

**Accommodation** - A modification to an assessment or its administration to allow access to the intended construct for a person with a disability. See **Construct, Disability**.

**Achievement Test** - A test that measures a specific body of knowledge or set of skills, usually after training or instruction has been received in the knowledge or skills. Tests of subject-matter competence in biology, calculus, literature, and so on are common examples of achievement tests. Compare **Aptitude Test**.

**Adaptive Test** - A test administered such that the items presented to a person depend, in part, on that person's response to a previous item or set of items. In general, correct responses lead to more difficult items, and incorrect responses lead to less difficult items. The goal is to focus on items that give the most information about an individual's level of ability. Adaptive tests are most often administered on a computer. Such tests are known as Computerized Adaptive Tests, or CATs.

**Adjusted Coefficient** - A statistic that has been revised to estimate its value under conditions other than those in the sample on which it has been calculated.

**Administration Mode** - The method by which an assessment is presented to the test taker, including, for example, printed booklets, Braille booklets, American Sign Language, computer display terminals, audio tapes, and videotapes. The most common administration mode has been paper booklets, but administration by means of computer is becoming increasingly common. See **Response Mode**.

**Algorithm** - A set of rules to be followed to accomplish some task. Often, algorithms are coded for implementation by computer.

**Alternate Forms** - Different editions of the same assessment, written to meet the same specifications and comparable in most respects, except that some or all of the questions are different. See **Specifications**.

## Glossary

**Alternate Form Reliability** - An estimate of reliability based on the correlation between alternate forms of an assessment administered to the same group of people. See **Alternate Forms, Reliability**. Compare **Test-Retest Reliability**.

**Analysis Sample** - The group of people on whose performance a statistic or set of statistics has been calculated.

**Anchor Test** - A short test or section administered with each of two or more forms of a test for the purpose of equating those forms. Compare **Common Items**. See **Equating**.

**Ancillary Materials** - Descriptive booklets, score interpretation guides, administration manuals, registration forms, etc., that accompany an assessment.

**Aptitude Test** - A test of general ability that is usually not closely related to a specific curriculum and that is used primarily to predict future performance. There is no strict distinction between aptitude tests and achievement tests. Compare **Achievement Test**.

**Assessment** - A systematic sample of behavior taken to allow inferences about an individual's knowledge, skill, ability, or other trait. Assessments could consist of multiple-choice tests, essay tests, portfolios, performance measures, structured interviews, etc. In most usages, "assessment" is synonymous with "test." Some authors use "assessment" to refer to broader evaluations than those provided by tests alone.

**Assessment Results** - Test scores or other indicators of test performance, including summaries of individual or group performance. See **Score**.

**Audit** - A systematic evaluation of a product or service with respect to documented standards to indicate whether or not the product or service is in compliance with the standards.

**Audit Model** - The methodology jointly agreed upon by an ETS program and the auditors for obtaining information about the audited product or service, evaluating the product or service, and documenting the results.

**Automated Item Selection** - A computer algorithm that follows a set of rules to assemble a test from a pool of items such that the test comes as close as possible to meeting the set of rules, given the constraints of the item pool.

**Automated Scoring of Constructed Responses** - A computer algorithm that follows a set of rules to generate a score for an essay or other constructed response. See **Constructed Response**.

**Bias** - In general usage, unfairness. In technical usage, a systematic error in estimating a parameter. See **Fairness**.

**Borderline Proficiency** - A level of knowledge, skill, or ability that is between the lowest level judged to be acceptable and the highest level judged to be unacceptable for some particular purpose. A cut score is often set in the area of borderline proficiency. See **Cut Score**.

**Branching Test** - An assessment in which test takers may be administered different sets of items, depending on their responses to earlier items or sets of items. See **Adaptive Test**.

**Canceled Score** - A canceled score is a score that has never been part of a test taker's record or is removed from a test taker's record. Such a score is not reportable. Scores may be canceled voluntarily by the test taker or by ETS for testing irregularities, misconduct, or score invalidity. See **Irregularity**.

**Certification** - Often, the granting of advanced status by a professional organization to an applicant who demonstrates appropriately high levels of skill and ability. Sometimes, certification is used as a synonym for licensing. Compare **Licensing**.

**Classification Error** - (1) The proportion of inconsistent or incorrect categorizations of test takers that would be made on repeated administrations of the same test or of a test and an alternate form, assuming no changes in the test takers' true score. (2) The assignment of a test taker to the wrong category, such as passing a person who lacks minimal competence and should fail. See **True Score**.

**Client** - An agency, organization, institution, individual, or the like that commissions ETS to provide a product or service.

**Coaching** - Short-term instruction in test-taking strategies and/or subject-matter review aimed directly at improving performance on a test.

**Common Items** - A set of test questions that remains the same in two or more forms of a test for purposes of equating. The common items may be dispersed among the items in the forms to be equated, or kept together as an anchor test. Compare **Anchor Test**. See **Equating**.

## Glossary

**Comparable Scores** - Scores that have been linked by some procedure. In some usages, comparable scores do not meet the criteria for being fully interchangeable. Compare **Equating**.

**Composite Score** - A score that is the combination of two or more scores by some specified formula.

**Computerized Adaptive Test (CAT)** - See **Adaptive Test**.

**Computer-Based Test** - Any test administered on a computer. Compare **Computerized Adaptive Test**.

**Construct** - The complete set of knowledge, skills, abilities, or traits an assessment is intended to measure, such as knowledge of American history, reading comprehension, study skills, writing ability, logical reasoning, honesty, intelligence, and so forth.

**Construct Label** - The name used to characterize the construct measured by a test. The construct label is generally not sufficient by itself to describe fully the set of knowledge, skills, abilities, or traits a test is intended to measure.

**Construct Relevant (Irrelevant) Variance** - Differences in scores among individuals related (unrelated) to differences in the knowledge, skills, abilities, or traits included in the intended construct. See **Construct**, **Variance**.

**Construct Validity** - All the theoretical and empirical evidence bearing on what an assessment is actually measuring, and on the qualities of the inferences made on the basis of the scores. Construct validity was previously associated primarily with assessments of abstract attributes, such as honesty, anxiety, or need for achievement, rather than assessments of more concrete attributes, such as knowledge of American history, ability to fly an airplane, or writing ability. Now construct validity is seen as the sum of all types of evidence bearing on the validity of any assessment. See **Validity**.

**Constructed Response** - An answer to a question or exercise generated by the test taker rather than selected from a list of possible responses.

**Content Validity** - The aspect of construct validity that emphasizes evidence bearing on the appropriateness of the knowledge, skills, and abilities measured by an assessment. Are the important areas of the domain represented by appropriate numbers of items? Are important areas of the domain excluded? Is material foreign to the domain included? See **Domain**, **Validity**.

**Conversion Parameters** - Quantitative rules for expressing scores on one assessment form in terms of scores on an alternate form. See **Alternate Forms, Equating**.

**Criterion** - That which is predicted by an assessment, such as college grade point average or job-performance rating.

**Criterion Referenced Test** - A test designed to be efficient in (1) determining the proportion of a domain the test taker has mastered, or in (2) determining whether or not the test taker has reached some necessary level of knowledge, skill, and ability. See **Absolute Score, Domain**. Compare **Norm Referenced Test**.

**Criterion-Related Validity** - The aspect of construct validity that emphasizes evidence bearing on the statistical relationships among assessment scores and other variables of interest. Often, the relationship is predictive. Criterion-related validity is usually expressed as a correlation coefficient. A common example of criterion-related validity evidence is the correlation between SAT scores and grades in college. See **Criterion, Validity**.

**Critical Content** - Knowledge, skills, abilities, or traits that must be measured in an assessment because of their importance in meeting the purpose of the assessment. For example, in an assessment used to license nuclear power plant operators, knowledge of safety procedures would be considered critical content.

**Cross Validation** - The application of scoring weights or prediction equations derived from one sample to a different sample, to allow estimation of the extent to which chance factors determined the weights or equations or inflated the validity estimated in the analysis sample.

**Customer** - A general term for those who sponsor, purchase, or use ETS products or services, including clients, institutional and individual score recipients, and test takers.

**Customer Satisfaction** - The extent to which customers feel they have been treated appropriately in their interactions with ETS, and the extent to which they feel ETS products and services are efficiently and effectively meeting their needs.

**Customer Service** - The extent to which interactions of ETS staff with customers increase customer satisfaction. See **Customer Satisfaction**.

## Glossary

**Cut Score** - A point on a score scale at or above which test takers are classified in one way and below which they are classified in a different way. For example, if a cut score is set at 60, then people who score 60 and above may be classified as "passing" and people who score 59 and below classified as "failing." See **Absolute Score, Standard**.

**Decision Consistency** - See **Reliability of Classification**.

**Difference Score** - The amount by which one score or subscore is higher (or lower) than another score or subscore. Difference scores tend to be unreliable. See **Reliability, Subscore**.

**Differential Item Functioning (DIF)** - An indication of differences in item performance among **comparable** members of different groups. Generally, in measures of DIF, individuals in different groups are considered comparable if they receive the same score on an assessment. Compare **Impact**.

**Disability** - A physical or mental impairment that substantially limits a major life activity. Individuals with disabilities may request non-standardized accommodations in order to have access to a standardized test. See **Accommodation**.

**Discrimination** - The power of an item to differentiate among test takers at different levels of ability on the construct being measured. In some usages, a synonym for bias. See **Bias**.

**Documentation** - Tangible evidence, generally in written form, of compliance with a standard, compiled for use in the audit process. See **Audit**.

**Domain** - A defined universe of knowledge, skills, abilities, attitudes, interests, or other characteristics.

**Equating** - A statistical process used to adjust scores on two or more alternate forms of an assessment so that the scores may be used interchangeably. See **Anchor Test, Common Items, Conversion Parameters**.

**Error** - In the context of assessment, nonsystematic fluctuations in scores caused by such factors as luck in guessing a response, the particular questions that happen to be in a form of an assessment, or whether the scorer is rigorous or lenient. Technically, error is the difference between an observed score and the true score for an individual. See **Observed Score, Standard Error of Measurement, True Score**.

**Error of Classification** - See **Classification Error**.

**ETS Board of Trustees** - The ETS Board of Trustees is the governing body of ETS. There are 17 trustees. Of these, 16 are elected for three-year terms. New members are elected by incumbent trustees. The president of ETS is an ex officio member.

**Face Validity** - Not really an aspect of validity, but simply test takers' impressions of the validity of an assessment based on the way it looks to them. See **Validity**.

**Fairness** - The extent to which a product or service is appropriate for members of different groups, and the extent to which users of products or services are treated the same way, regardless of gender, race, ethnicity, and the like. For assessments, there are many, often conflicting, definitions of fairness. Some definitions focus on equal outcomes for people with the same scores, regardless of group membership. Other definitions focus on equal outcomes for different groups, even if the people have different scores. One useful definition of fairness is that an assessment is fair if any group differences in performance are valid. The existence of group differences in performance does not necessarily make an assessment unfair, because the groups may really differ on the construct being measured. See **Validity**.

**Flag** - A notation that may accompany a score under certain conditions to indicate that the assessment and/or the conditions of administration were modified in ways that may have changed the meaning of the score.

**Form** - The set of assessments all assembled to the same specifications with the same questions.

**Formative Evaluation** - Appraisals of a product or service as it is being designed and developed to help ensure that the final version is appropriate. Compare **Summative Evaluation**.

**Formula Score** - Raw score on a multiple-choice test after a correction for guessing has been applied, usually the number of right responses minus a fraction of the number of wrong responses. Compare **Raw Score**.

**Grade Equivalent Score** - A score expressed as the school year and month for which the reported performance is average. A grade equivalent score of 3.6, for example, means the test taker performed as well as the average student in the norm group who was in the sixth month of the third grade. See **Norm Group**.

**Impact** - A raw difference between groups in percent correct on an item, or in scores or passing rates on an assessment. Compare **Differential Item Functioning**.

## Glossary

**Imputed Value** - An imputed value is an estimate that is used in place of the unknown value of an observable variable.

**Information Curve** - The graph of a mathematical function showing the precision with which test takers' abilities are estimated at different points on the ability scale. Easy assessments generally give more information at lower levels of ability and difficult assessments generally give more information at higher levels of ability.

**Intended Population** - The people for whom an assessment has been designed to be most appropriate.

**Irregularity** - Problem, disruption, or unacceptable behavior at an assessment administration.

**Item** - An assessment question or task of any type.

**Item Analysis** - A statistical description of how an item performed within a particular assessment when administered to a particular sample of people. Data often provided are the difficulty of the question, the number of people choosing each of the options, and the correlation of the item with the total score or some other criterion.

**Item Response** - (1) A person's answer to a question. (2) The answer to a question coded into categories such as right, wrong, or omit.

**Item Response Theory (IRT)** - A mathematical model relating performance on questions (items) to certain characteristics of the test takers and certain characteristics of the items.

**Item Type** - The observable format of a question. At a very general level, "item type" may, for example, refer to multiple-choice or free-response questions. At a finer level of distinction, "item type" may, for example, refer to synonym questions or antonym questions.

**Joint Standards** - See *Standards for Educational and Psychological Testing*.

**Key** - A correct answer to a question, or a listing of the correct responses to a set of assessment questions.

**Licensing** - The granting by a government agency of permission to practice an occupation or profession based on evidence that the applicant has at least the minimally acceptable level of knowledge and skills needed to protect the public from harm. Compare **Certification**.

**Linguistic Demands** - The reading or listening ability necessary to comprehend an assessment's questions or tasks, and the writing or speaking ability necessary to respond.

**Linking** - The general term for making scores on assessments more or less comparable to one another. It can range from strict statistical equating that results in scores that are interchangeable, to social moderation based on the subjective judgments of some group of people. Compare **Equating**.

**Linking Items** - See **Common Items**.

**Local Norms** - A distribution of scores and related statistics within an institution or closely related group of institutions (such as the schools in one district) used to give additional meaning to scores by serving as a basis for comparison. See **Norms**.

**Locally Administered Assessment** - An assessment that is given by an institution at a time and place of the institution's own choosing.

**Mastery Test** - See **Criterion Referenced Test (2)**.

**Matrix Sampling** - A method of assessment administration in which different samples of students are given different (possibly overlapping) samples of items. Matrix sampling is an efficient means of gathering data for groups because no individual has to respond to all of the items that are administered.

**Mean** - The arithmetic average. The sum of a set of scores divided by the number of scores.

**Meta Analysis** - A method of combining the results of a number of studies to gain statistical power.

**Misuse** - The use of an assessment in ways other than those intended, resulting in harmful effects.

**Modification** - See **Accommodation**.

**Multi-factorial** - Measuring several different knowledge areas, skills, or abilities.

**Multiple-Choice Test** - An assessment in which the test taker selects the correct response to an item from a limited number of answer choices (generally four or five).

## Glossary

**Multiple Prediction** - The use of more than one predictor for the same criterion. For example, SAT scores and high school grade point average may be used together to predict college grade point average. See **Criterion**.

**Norm Group** - A sample of test takers, usually representative of some relevant population, for whom performance data have been obtained. The scores of test takers are given meaning by comparison to the distribution of scores in the norm group. See **Normative Scale, Norms**.

**Norm Referenced Test** - A test designed to allow the scores of test takers to be compared with the scores of one or more norm groups. See **Norm Group, Normative Scale, Norms**. Compare **Criterion Referenced Test**.

**Normative Scale** - A way of expressing a score's relative standing in the distribution of scores of some specified group. A common normative scale, for example, is percentile rank in which a score is expressed as the percent of the norm group scoring at that level. See **Norm Group, Norms, Percentile**.

**Norms** - Performance data for a sample from a relevant population (norm group) used to add meaning to test takers' scores. For example, saying that someone answered 36 items correctly does not carry much information. Saying that the score of 36 is at the 84th percentile for a national sample of third-grade students adds meaning to the score. See **Normative Scale, Norm Group, Percentile**.

**Observed Score** - The score a person happens to obtain on a particular form of an assessment at a particular administration. Compare **True Score**.

**Operational Administration** - The use of an assessment to obtain scores that will be used for their intended purposes. Compare **Pretest**.

**Parameter** - (1) The value of some variable for a population as distinguished from an estimate of the value based on a sample drawn from the population. (2) In item response theory, one of the characteristics of an item, such as its difficulty. See also **Conversion Parameter**.

**Pass-Fail Score** - See **Cut Score, Performance Standard**.

**Percent Correct Score** - The number of questions answered correctly divided by the number of questions on the assessment. Percent correct scores are not comparable across different assessments. For example, 50 percent correct on a difficult measure may indicate a higher level of knowledge than 80 percent correct on an easy measure. Compare **Percentile Rank**.

**Percentile Rank** - The percent of some defined group who scored below (in some usages, at or below) a particular score on an assessment. For example, a score at the 84th percentile means that 84 percent of some group obtained lower scores. Compare **Percent Correct Score**.

**Performance Test** - An assessment in which test takers actually do relatively realistic tasks rather than answer questions, such as teach a class, parallel park a car, play a particular piece of music, complete a chemistry experiment, repair a clogged fuel injector, perform an appendectomy, land an airplane, or use some software package.

**Pilot Testing** - A form of pretest consisting of a small-scale tryout of questions or an assessment form, often involving observation of and interviews with test takers. Compare **Pretest, Operational Administration**.

**Pool** - The set of available items from which an assessment or group of assessments will be assembled.

**Population** - All the members of some defined group, such as third-grade students in the United States. Generally, populations are too large to be assessed and smaller samples are drawn. Compare **Sample**.

**Population Group** - A part of a larger population that is defined by various criteria such as gender, race or ethnic origin, training or formal preparation, geographic location, income level, disability, or age.

**Portfolio** - A systematic collection of materials demonstrating a person's level of knowledge, skill, or ability in a particular area. For example, portfolios may consist of a collection of essays written at different times on different topics to show writing ability, or a collection of lesson plans, videotaped lessons, and written self-evaluations to show teaching ability.

## Glossary

**Precision** - The width of the interval within which a value can be estimated to lie with a given probability. The higher the precision, the smaller the interval required to include the value at any given probability. For assessment scores, precision has the same meaning as reliability. See **Reliability**.

**Predictive Validity** - See **Criterion-Related Validity**.

**Preliminary Item Analysis** - An item analysis performed after an operational assessment has been administered, but before scores are released. It is used as a quality control measure. See **Item Analysis**.

**Presentation Mode** - See **Administration Mode**.

**Pretest** - A nonoperational trial administration of items or an assessment to gather data on item or assessment characteristics, such as difficulty and discrimination. See **Discrimination**. Compare **Operational Administration**.

**Profile Scores** - A number of scores or subscores interpreted with respect to their relative magnitudes. For example, a test taker who scores higher in mathematics than in reading would have a profile different from one who scored higher in reading than in mathematics.

**Program** - An integrated group of ETS products or services serving similar purposes and/or similar populations. A program is characterized by its continuing character and by the inclusiveness of the services provided.

**Program Statistics** - Data based on the people who happen to have taken a particular assessment during some specified interval. Because program statistics are generally based on self-selected samples, generalization to larger populations is not appropriate. For example, it is wrong to make inferences about all high school seniors from the high school seniors who choose to take the SAT.

**Raw Score** - (1) The number of items answered correctly. (2) In some usages, the formula score is also called a raw score. See **Formula Score**.

**Registration** - The process of enrolling to take an ETS assessment.

**Regression Equation** - A formula, often of the form  $Y = aX + b$ , used to estimate the value of a criterion, given the value of one or more observed variables used as predictors. For example, a regression equation is used to estimate college grade point average, given high school grade point average and SAT scores. See **Criterion**.

**Relative Score** - A score that takes on meaning by comparison with a distribution of scores, such as percentile rank or grade equivalent scores. Compare **Absolute Score**. See **Grade Equivalent Score**, **Percentile Rank**.

**Reliability** - An indicator of the extent to which scores will be consistent across different administrations, and/or administration of alternate forms of the assessment, and/or different scorers. Reliability is also defined as the ratio of true score variance to total score variance. See **Alternate Form Reliability**, **Test-Retest Reliability**, **True Score**, **Variance**.

**Reliability of Classification** - An indicator of the extent to which assessment scores will be consistent in assigning a test taker to some category, such as pass or fail, master or non-master, if the test taker is retested with the same assessment or an alternate form of the assessment, assuming no relevant change in the test taker. See **Alternate Form**.

**Replicate** - To repeat a study to determine whether the results are consistent.

**Response Mode** - The procedure used by a test taker to indicate an answer to a question, such as a mark on an answer sheet, a handwritten essay, or an entry on a computer keyboard. Compare **Administration Mode**.

**Restriction of Range** - A case in which the variance in a sample is lower than the variance in the population. See **Sample**, **Population**, **Variance**.

**Sample** - A subset of a larger population. For example, a few hundred high schools may be selected to represent the more than 20,000 high schools in the United States. Samples differ in how well they represent the larger population. Generally, the care with which a sample is chosen has a greater effect on its ability to represent a population than does the size of the sample.

**Sampling Error** - The difference between a statistic derived from a particular sample and the corresponding parameter for the population from which the sample was drawn. See **Parameter** (1), **Population**, **Sample**.

**Score** - A quantitative or categorical value (such as "pass" or "fail") assigned to a test taker as the result of some measurement procedure.

## Glossary

**Score Recipient** - A person or institution obtaining the scores of individual test takers or summary data for groups of test takers.

**Score Scale** - The set of numbers within which scores are reported for a particular assessment or program, often, but not necessarily, having a specified mean and standard deviation for some defined reference group. For example, when the SAT was recentered, the mean was set at 500, and the standard deviation was set at 100 on a scale that runs from 200 to 800. See **Mean, Standard Deviation**.

**Scoring Rubric** - A set of rules and guidelines for assigning scores to free-response or performance items. Generally, there is a description of the attributes of responses associated with each score level. Often rubrics are accompanied by examples of responses at various score levels.

**Service Standards** - Criteria for evaluating interactions with customers, such as average number of rings before an 800 number is answered, hours when phones will be answered, or speed of order fulfillment.

**Specifications** - Detailed documentation of the intended characteristics of an assessment, including but not limited to the content and skills to be measured, the numbers and types of items, the level of difficulty and discrimination, the timing, and the layout.

**Speededness** - The extent to which test takers lack sufficient time to respond to items. Some assessments are speeded on purpose, if speed of response is an aspect of the construct being measured. For most assessments, however, speededness is not a desirable characteristic. See **Construct**.

**Sponsors** - Educational, professional, or occupational associations; federal, state, or local agencies; or public or private foundations that contract with ETS for its products or services. See **Client**.

**Standard** - (1) A cut score or a defined minimally acceptable level of performance on some task. For example, answer 80 out of 100 items correctly, or run 100 yards in 12 seconds or less. In some usages, a description of a desired level of performance. See **Cut Score**.  
(2) A ruling guide or principle.

**Standard Deviation** - A statistic characterizing the magnitude of the differences among a set of scores. The more spread out the scores, the larger the standard deviation. Specifically it is the square root of the average squared difference between each score and the mean of the scores. The standard deviation is the square root of the variance. See **Variance**.

**Standard Error of Estimate** - A statistic that indicates the standard deviation of differences between actual and predicted scores. See **Standard Deviation**.

**Standard Error of Measurement** - In general terms, a statistic that indicates the "wobble" expected in assessment scores because of random differences caused by such factors as luck in guessing a response, the particular set of items in the form administered to the test taker, or the leniency or rigor of a scorer. In more technical terms, the standard error of measurement is a statistic that indicates the standard deviation of the differences between observed scores and their corresponding true scores. It is also, theoretically, the standard deviation of scores for a person taking a large number of parallel forms of assessments, assuming no changes in the person's true score. See **Error**, **True Score**, **Standard Deviation**.

**Standardized Conditions** - The administration of an assessment in the same manner to all test takers to allow fair comparison of their scores. Factors such as timing, directions, and use of aids (e.g., calculators) are controlled to be constant for all test takers.

*Standards for Educational and Psychological Testing* - A document published by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). It lists what the publishers purport to be the appropriate ways to develop, use, and evaluate assessments.

**Studied Group** - A population group sampled in evaluations of the performance of an assessment for people in different groups, particularly with respect to fairness. See **Population Group**.

**Subscore** - A score derived from a subset of the items in an assessment.

**Subtest** - A subset of the items in an assessment upon which a subscore or part score is based. See **Subscore**.

**Summative Evaluation** - Appraisals of a product or service after it has been completed, to help determine whether or not the final version is appropriate. Compare **Formative Evaluation**.

**Target Information Curve** - The desired information curve for an assessment indicating the parts of the ability scale where the most precise measurement is required. See **Information Curve**.

## Glossary

**Test** - See **Assessment**.

**Test Analysis** - A description of the statistical characteristics of an assessment following administration, including but not limited to distributions of item-difficulty and discrimination indices, score distributions, mean and standard deviation of scores, reliability, standard error of measurement, and indices of speededness.

**Test Battery** - A collection of tests often administered together.

**Test Center** - A site where assessments are administered.

**Test Edition** - See **Form**. Compare **Alternate Forms**.

**Test Format** - (1) The physical layout of a test, including the spacing of items on a page, type size, positioning of item-response options, and so forth. (2) Also used to refer to the Administration Mode and Response Mode. See **Administration Mode**, **Response Mode**.

**Test-Retest Reliability** - An estimate of reliability based on the correlation between scores on two administrations of the same form to the same group of people. Because the items remain unchanged, they are not counted as a source of error variance in the estimate of reliability. See **Error**, **Reliability**, **Variance**. Compare **Alternate Form Reliability**.

**Timeliness** - The degree to which a product or service is delivered to its recipient within a predefined schedule.

**True Score** - The hypothetical average score of a test taker calculated from an infinite number of administrations of alternate forms, assuming no changes in learning, forgetting, or fatigue on the part of the test taker. It is the score that a test taker would obtain if the assessment were perfectly reliable (the standard error of measurement were zero). See **Alternate Form**, **Reliability**, **Standard Error of Measurement**. Compare **Observed Score**.

**User** - Individual or institution making decisions on the basis of assessment scores.

**Validity** - The extent to which inferences and actions made on the basis of a set of scores are appropriate and justified by evidence. It is the most important aspect of the quality of an assessment. Validity refers to how the scores are used rather than to the assessment itself. Validity is a unified concept, but several aspects of validity evidence are often distinguished. Compare **Construct Validity**, **Content Validity**, **Criterion-Related Validity**, **Face Validity**.

**Validity Argument** - A coherent and rational compilation of evidence designed to convey all of the information available to a program concerning the validity of an assessment's scores for a particular purpose. See **Validity**.

**Variable** - An attribute that can take on different values, such as scores, grade point average, family income, age, and weight.

**Variance** - (1) Generally, a label for differences among scores. "Sources of variance," for example, refers to causes of the differences among test takers' scores. (2) Technically, a statistic characterizing the magnitude of the differences among a set of measurements. Specifically, it is the average squared difference between each measurement and the mean of the measurements. The square root of the variance is the Standard Deviation. See **Standard Deviation**.

**Weighting** - (1) A formula giving the relative contribution of part scores or items to a composite score. See **Composite Score**. (2) The relative contribution assigned to certain sample data to represent a population more accurately. See **Population, Sample**.



01603-004663 • U112E10  
I.N. 996171