Listening. Learning. Leading.®

ETS Standards

for Quality and Fairness

2014

2014

ETS Standards

for Quality and Fairness

ETS®

Listening. Learning. Leading.®

Table of Contents

Preface 1

Introduction 2

Purpose 2

Relation to Previous *ETS Standards* and to the *Standards for Educational and Psychological Testing* 2

Application of *ETS Standards* 2

Audit Program 3

Overview 4

CHAPTER 1 **Corporate Responsibilities** 5

Purpose 5

Standards 1.1–1.11 5

CHAPTER 2 **Widely Applicable Standards** 9

Purpose 9

Standards 2.1–2.5 9

CHAPTER 3 **Non-Test Products and Services** 11

Purpose 11

Standards 3.1–3.7 11

CHAPTER 4 **Validity** 15

Purpose 15

Standards 4.1–4.7 15

CHAPTER 5 **Fairness** 19

Purpose 19

Standards 5.1–5.7 19

CHAPTER 6 **Reliability** 25

Purpose 25

Standards 6.1–6.6 25

CHAPTER 7 **Test Design and Development** 29

Purpose 29

Standards 7.1–7.8 29

CHAPTER 8 **Equating, Linking, Norming, and Cut Scores** 35

Purpose 35

Standards 8.1–8.10 35

CHAPTER 9 **Test Administration** 39
Purpose 39
Standards 9.1–9.6 39

CHAPTER 10 **Scoring** 43
Purpose 43
Standards 10.1–10.4 43

CHAPTER 11 **Reporting Test Results** 45
Purpose 45
Standards 11.1–11.5 45

CHAPTER 12 **Test Use** 49
Purpose 49
Standards 12.1–12.5 49

CHAPTER 13 **Test Takers' Rights and Responsibilities** 51
Purpose 51
Standards 13.1–13.5 51

Glossary 54

Preface

The *ETS Standards for Quality and Fairness* are central to our mission to advance quality and equity in education for learners worldwide. The *ETS Standards* provide benchmarks of excellence and are used by ETS staff throughout the process of design, development, and delivery to provide technically fair, valid, and reliable tests, research, and related products and services. Program auditors use the *ETS Standards* in a thorough internal audit process to evaluate our products and services according to these established benchmarks. The ETS Board of Trustees oversees the results of the audit process to ensure successful implementation of the *ETS Standards*.

The *ETS Standards* were initially adopted as corporate policy by the ETS Board of Trustees in 1981. They are periodically revised to ensure alignment with current measurement industry standards as reflected by the *Standards for Educational and Psychological Testing*, published jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. This edition of the *ETS Standards* also reflects changes in educational technology, testing, and policy, including the emphasis on accountability, the use of computer-based measurement, and on testing as it relates to English-language learners and individuals with disabilities.

The *ETS Standards for Quality and Fairness* and audit process help to ensure that we provide tests, products, and services that reflect the highest levels of quality and integrity, and that we deliver tests and products that meet or exceed current measurement industry standards. They help us achieve the mission, and demonstrate our commitment to public accountability. The *ETS Standards* and audit process are a model for organizations throughout the world that seek to implement measurement standards aligned with changes in technology and advances in measurement and education.

Walt MacDonald
President and CEO
Educational Testing Service

Introduction

Purpose

The purposes of the *ETS Standards for Quality and Fairness* (henceforth the *SQF*) are to help Educational Testing Service design, develop, and deliver technically sound, fair, accessible, and useful products and services, and to help auditors evaluate those products and services. Additionally, the *SQF* is a publicly available document to help current and prospective clients, test takers, policymakers, score users, collaborating organizations, and others understand the requirements for the quality and fairness of ETS products and services.

The *SQF* is designed to provide policy-level guidance to ETS staff. The individual standards within the document are put into practice through the use of detailed guidelines, standard operating procedures, work rules, checklists, and so forth.

Relation to Previous *ETS Standards* and to the *Standards for Educational and Psychological Testing*

This edition of the *SQF* owes much to earlier versions of the document as first adopted by the ETS Board of Trustees in 1981 and as updated in 1987, 2000, and 2002. The earlier versions of the *SQF* and the accompanying audit process stand as tangible evidence of the long-standing willingness of ETS to be held accountable for the quality and fairness of its products and services.

ETS strives to follow the relevant standards in the 2014 *Standards for Educational and Psychological Testing* (also called the *Joint Standards*) issued by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The *SQF* is intended to be consistent with the *Joint Standards*, but the contents have been tailored to the specific needs of ETS. The *Joint Standards* is intentionally redundant, with the expectation that readers will focus only on certain chapters. The *SQF* is far less redundant, with the expectation that users will become familiar with all of the chapters relevant to their work. The *SQF* includes some material not included in the *Joint Standards* (e.g., non-test products and services), and excludes some material found in the *Joint Standards* that is not applicable to ETS products or services (e.g., clinical aspects of testing). Furthermore, the *SQF* is intended for use by ETS staff, and does not directly address others involved in testing such as designers of policy studies, program evaluators, and state and district testing directors, as does the *Joint Standards*.

Application of *ETS Standards*

The application of the *ETS Standards* in the *SQF* 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 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.

ETS does not always control all aspects of a product or service to which ETS staff contribute. Collaboration with other organizations has become common. Whenever possible, adherence to the *SQF* should be part of collaborative agreements, but ETS cannot force others who have independent control of parts of a product or service to comply with the *SQF*.

Audit Program

The audit program established to monitor compliance with the original *SQF* will continue to do so with the 2014 version. The purpose of the ETS Audit Program is to help ensure that products and services provided by ETS will be evaluated with respect to rigorous criteria, using a well-documented process. Those products and services should be periodically audited for compliance with the *SQF* in an effort to ensure their quality and fairness.

The ETS Office of Professional Standards Compliance (OPSC) establishes the audit schedules to ensure that ETS products and services are audited at reasonable intervals, generally once every three years. In consultation with the ETS Office of the General Counsel, the OPSC may extend the regularly scheduled audit cycle based on excellent results in previous audits for products or services that are essentially unchanged since their last audit, or for other reasons that the OPSC deems sufficient.

The OPSC recruits auditors to perform each review. Auditors reflect the diversity of ETS professional staff. The auditors assigned to a product or service are independent of the product or service being audited and, as a group, have the knowledge and experience necessary to make the required judgments about the product or service being evaluated.

The OPSC organizes audit teams to perform the reviews. In addition to members of ETS staff, individuals from outside ETS serve as members of some audit teams to provide fresh insights and public perspectives. The OPSC trains auditors and program staff to perform their roles in the audit process.

Program staff members evaluate the compliance of their products and services with each of the relevant 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 a process agreed upon by the program, the auditors, and the OPSC. Whenever members of an audit team believe that a product or service does not comply with a relevant standard, they must explain why and make an appropriate recommendation for resolving the situation.

Participants in each audit work together to facilitate a thorough and efficient review, in consultation with staff in the OPSC, and clients as appropriate. Programs, possibly in collaboration with clients, develop and implement action plans as necessary to bring their product or service into compliance with the *SQF* as promptly as possible. A corporate-level Ratings Panel reviews all audit results, including action plans, and determines a holistic rating of each program's compliance with the *SQF*.

The OPSC monitors progress in bringing a program into compliance with the *SQF* and 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 integrity of the entire process.

Overview

There are 13 chapters and a glossary following this introduction:

- Chapter 1: Corporate Responsibilities
- Chapter 2: Widely Applicable Standards
- Chapter 3: Non-Test Products and Services
- Chapter 4: Validity
- Chapter 5: Fairness
- Chapter 6: Reliability
- Chapter 7: Test Design and Development
- Chapter 8: Equating, Linking, Norming, and Cut Scores
- Chapter 9: Test Administration
- Chapter 10: Scoring
- Chapter 11: Reporting Test Results
- Chapter 12: Test Use
- Chapter 13: Test-Takers' Rights and Responsibilities

The chapters titled "Corporate Responsibilities," "Widely Applicable Standards," and "Scoring" are new to the 2014 *SQF*. Chapters 1, 2, and 5 apply to all ETS products, services, and activities. Chapter 3 applies to all ETS products and services except tests. All of the other chapters apply to tests and test-related activities. The standards that apply to tests are relevant for all types of tests regardless of format or construct measured. In addition to traditional multiple-choice and constructed-response tests, the standards apply to formative tests, games-based tests, questionnaires, noncognitive measures, portfolios, and any other form of evaluation developed by ETS, as long as decisions are made based on the results.

The division into separate chapters may be misleading in certain respects. Fairness, for example, is a pervasive concern, and standards related to fairness could appropriately occur in many chapters. Placing most of the fairness-related standards in a single chapter is not meant to imply that they are isolated from other aspects of testing.

Some of the placement of standards into chapters is somewhat arbitrary. A standard on fairness in scoring, for example, is relevant to both the "Fairness" chapter and the "Scoring" chapter. In an effort to avoid redundancy, it is placed in only one of the chapters. Therefore, the various chapters are NOT independent and cannot stand alone. ETS staff and external auditors who use the *SQF* are expected to become familiar with all of the chapters related to their work.

CHAPTER 1

Corporate Responsibilities

Purpose

The purpose of this chapter is to state the corporate standards that apply to all ETS activities and to all users of the *ETS Standards*.

The standards focus on the need for all ETS programs and services to support the ETS mission, to operate within applicable laws, to ascertain and meet the needs of customers, to maintain records, and to be accountable for the utility and quality of ETS products and services.

Standards

Standard 1.1: Conforming with the ETS Mission

Every ETS product or service must be in conformity with the ETS mission to help advance quality and equity in education by providing fair and valid tests, research, and related services.

Indicate how each product, service, or major activity contributes to the ETS mission. Products and services that meet the ETS mission must be suitable for their intended purpose, must be technically sound, and must be appropriate for diverse groups within the intended population of users. Avoid products, services, or activities that are contrary to the ETS mission.

Standard 1.2: Complying with Laws

All programs and activities must comply with applicable laws and regulations.

Consult the ETS Office of the General Counsel as necessary to help ensure that all ETS programs, activities, and operations are legally compliant.

Standard 1.3: Using Resources Appropriately

Use ETS funds and ETS property only for their intended purposes.

Program staff should use ETS resources appropriately.

Standard 1.4: Protecting Privacy and Intellectual Property

Protect the privacy of test takers and research subjects, the security of personally identifiable information, the intellectual property rights of ETS and its clients, and the security of confidential materials.

Follow appropriate procedures to maintain the privacy of test takers and research subjects. Protect ETS's and the client's rights with respect to such proprietary products as confidential test items, software, marketing studies, procedural manuals, trade secrets, new product development plans, trademarks, copyrights, and the like. Develop, document, and follow procedures for maintaining the security of confidential materials in all media (electronic or print) to reduce the likelihood of unauthorized disclosure, to the extent feasible.

Standard 1.5: Making Information Available

Provide convenient methods for members of the public, customers, and other interested parties to obtain information, ask questions, make comments, or register problems or concerns. If an answer is required, respond promptly and courteously. Upon request, provide reasonable access to ETS-controlled, nonproprietary information about ETS, about ETS products and services, and about research studies and results, within the constraints established in Standard 1.4.

The default position should be to respond positively to reasonable requests for information whenever it is appropriate to do so. It is particularly important to grant access to data facilitating the reanalysis and critique of published ETS research.

Standard 1.6: Retaining Information and Records

Retain the information and records necessary to verify reported scores, research results, and program finances.

Establish and follow data and records retention policies approved by the Office of the General Counsel including guidelines for the destruction of files that no longer need to be retained.

Standard 1.7: Maintaining Continuity of Operations

Establish business continuity and disaster recovery plans, and implement procedures to protect crucial information and maintain essential work processes in the event of a disaster.

Back up essential systems and data in safe locations not likely to be affected by disasters at the primary location for data processing and data storage.

Standard 1.8: Obtaining Customer Input

Identify the customers of a product or service, and obtain their input into the design, development, and operation of the product or service.

The term "customer" includes clients, users, and purchasers of ETS products or services. For testing programs, "customer" also includes test takers and users of scores. For products or services developed for a particular client, work collaboratively with the client as appropriate during the design, development, operation, and evaluation of the products or services.

Standard 1.9: Evaluating Service Quality

Develop, in consultation with clients as appropriate, the measures by which service quality will be evaluated. Include such variables as response times for inquiries, speed of order fulfillment, and accuracy and timeliness of deliverables.

Document the quality measures and agreed-upon service levels. Monitor and document the extent to which the agreed-upon service levels are met.

Standard 1.10: Measuring Customer Satisfaction

Periodically measure customer satisfaction. As appropriate, use the information to improve levels of customer satisfaction.

Obtain information from customers concerning their satisfaction with products and services and their interactions with ETS. 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, web comments, client interviews, customer advisory groups, market research studies, process metrics, reviews of customer complaints, and so forth.

Standard 1.11: Preventing Problems and Verifying Accuracy

Prevent problems and address risks in all phases of planning, developing, and delivering products and services. Verify the accuracy of deliverables to customers. Correct any errors that will affect the achievement of agreed-upon customer expectations. Make the corrections in a timely manner that is responsive to customer needs. Document problems that affect customers and use the information to help avoid future problems and risks.

Design the process to reduce the likelihood of problems and to provide for the early detection of those problems it is impossible to avoid. Monitor the progress of work against schedules using sound project management methods. Notify customers likely to be adversely affected if agreed-upon important deadlines for deliverables will not be met. Products reaching customers should, to the extent possible, adhere to specifications, service-level agreements, and agreed-upon customer expectations.

Contact the Office of Quality for information about appropriate methods for the prevention and resolution of problems.

CHAPTER 2

Widely Applicable Standards

Purpose

The purpose of this chapter is to avoid needless redundancy by stating in a single place those widely applicable standards that otherwise would have to be repeated in many chapters.

For example, all ETS communications should be clear, accurate, and understandable by the intended recipients whether concerning reliability, validity, fairness, rules for administering tests, test scores, test takers' rights, or anything else. Rather than repeating a standard about communicating well in each applicable chapter, that standard, and others of similar generality, are stated in this chapter.

Standards

Standard 2.1: Communicating Accurately and Clearly

All communications (e.g., advertisements, press releases, proposals, directions to test takers and test administrators, scoring rubrics, score reports, information for score users) should be technically accurate and understandable by the intended receivers of such communications.

No communication should misrepresent a product or service or intentionally mislead the recipient of the communication. Logical and/or empirical support should be available for the claims ETS makes about any ETS product or service.

Express any statistical information for score users (e.g., reliability statistics) in terms that the score users can reasonably be assumed to understand. Make sure that any accompanying explanations are both technically correct and understandable to the intended recipients. Avoid using terms that the intended recipients are likely to misunderstand (e.g., "measurement error"). If it is not possible to avoid the use of such a term, explain it in language that the intended recipients are likely to understand.

If data or the results of research are reported, provide information to help the intended recipients interpret the information correctly. If the original research included important information about how the results should be interpreted, later communications about the results should include or refer to the information. If statistics are reported, indicate the degree of uncertainty associated with them. If adjusted statistics are reported, such as for restriction of range, make clear that the reported statistics have been adjusted and either report the unadjusted statistics or indicate where they may be obtained.

Standard 2.2: Documenting Decisions

Document the important decisions made about a product or service as it is designed, developed, and used. Include a rationale for each of the decisions in the documentation.

Keep a record of the major decisions (e.g., construct to be measured by a test, population to be sampled in a research study, equating method to be used by a testing program) affecting products and services, the people who made those decisions, and the rationales and data (if any) supporting the decisions. Make the information available for review during the audit process.

Standard 2.3: Using Qualified People

The employees or external consultants assigned to a task should be qualified to perform the task. Document the qualifications (e.g., education, training, experience, accomplishments) of the people assigned to a task.

For example, item writers and reviewers should have both subject-matter knowledge and technical knowledge about item writing. Fairness reviewers should have training in the identification of symbols, language, and content that are generally regarded as sexist, racist, or offensive. The psychometricians who design equating studies should be knowledgeable about gathering the necessary data and selecting and applying the appropriate equating model, and so forth.

Standard 2.4: Using Judgments

If decisions are made on the basis of the judgments or opinions of a group of people, such as subject-matter experts (e.g., developing test specifications, evaluating test items, setting cut scores, scoring essays), describe the reason for using the people, the procedures for selecting the people, the relevant characteristics of the people, the means by which their opinions were obtained, the training they received, the extent to which the judgments were independent, and the level of agreement reached.

The relevant characteristics of the people include their individual qualifications to be judges and, if the data are available, the extent to which the demographic characteristics of the group of people represent the diversity of the population from which they were selected. Specify the definition of "agreement" among judges if it is other than exact agreement.

Standard 2.5: Sampling

If a sample is drawn from a population (e.g., items from a pool, people from a group of test takers, scores from a distribution, schools from a state), describe the sampling methodology and any aspects of the sample that could reasonably be expected to influence the interpretation of the results.

Indicate the extent to which the sample is representative of the relevant population. Point out any material differences between the sample and the population, such as the use of volunteers rather than randomly selected participants, or the oversampling of certain subgroups.

CHAPTER 3

Non-Test Products and Services

Purpose

The purpose of this chapter is to help ensure that non-test products and services are capable of meeting their intended purposes for their intended populations, and will be developed and revised using planned, documented processes that include advice from diverse people, reviews of intermediate and final products, and attention to fairness and to meeting the needs of clients and users.

This chapter applies to non-test products and services intended for use outside of ETS. Products and services should be developed and maintained through procedures that are designed to ensure an appropriate level of quality. ETS products and services should be designed to satisfy customers' needs.

There should be documented logical and/or empirical evidence that the product or service should perform as intended for the appropriate populations. The evidence could include such factors as the qualifications of the designers, developers, and reviewers of the product or service; the results of evaluations of aspects of the product or service in prototypes or pilot versions; and/or the opinions of subject-matter experts. Evidence based on controlled experiments or usability studies is welcomed, but is not required.

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.

It is not the intent of this chapter to require a single development procedure for all ETS non-test products and services.

Standards

Standard 3.1: Describing Purpose and Users

Describe the intended purposes of the product or service and its desired major characteristics. Describe the intended users of the product or service and the needs that the product or service meets.

Provide sufficient detail to allow reviewers to evaluate the desired characteristics in light of the purposes of the product or service. The utility of products and services depends on how well the needs of the intended users are identified and met.

Standard 3.2: Establishing Quality and Fairness

Document and follow procedures designed to establish and maintain the technical quality, utility, and fairness of the product or service. For new products or services or for major revisions of existing ones, provide and follow a plan for establishing quality and fairness.

Obtain logical and/or empirical evidence to demonstrate the technical quality, utility, and fairness of a product or service.

Standard 3.3: Obtaining Advice and Reviews

Obtain substantive advice and reviews from diverse internal and external sources, including clients and users, as appropriate. Evaluate the product or service at reasonable intervals. Make revisions and improvements as appropriate.

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 customers 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. The program should determine the appropriate interval between periodic reviews and provide a rationale for the time selected.

Standard 3.4: Reassessing Evidence

If relevant factors change, reassess the evidence that the product or service meets its intended purposes for the intended populations, and gather new evidence as necessary.

Relevant factors include, but are not limited to, substantive changes in intended purpose, major changes to the product or service itself, or the way it is commonly used, and changes in the characteristics of the user population.

Standard 3.5: Providing Information

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. Inform users of the product or service how to gather evidence that the product or service is meeting its intended purpose.

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

If the product is available in varied formats such as computer-based and print-based, provide information about the characteristics of the formats and the relative advantages and disadvantages of each. Provide advice upon request concerning how to run local studies of the effectiveness of the product or service.

Standard 3.6: Warning Against Misuse

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

No program can anticipate or warn against 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.

Standard 3.7: Performing Research

ETS research should be of high scientific quality and follow established ethical procedures. Obtain reviews of research plans to help ensure the research is worthwhile and well designed. Obtain informed consent from human subjects (or the parents or guardians of minor subjects) as necessary. Minimize negative consequences of participation in research studies to the extent possible. 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, the soundness of the design, the thoroughness of data collection, the appropriateness of the analyses, and the fairness of the report. If the purpose of a test is for research only and operational use is not intended, the test materials should indicate the limited use of the test.

CHAPTER 4

Validity

Purpose

The purpose of this chapter is to help ensure that programs will gather and document appropriate evidence to support the intended inferences from reported test results and actions based on those inferences.

Validity is one of the most important attributes of test quality. Programs should provide evidence to show that each test is capable of meeting its intended purposes.

Validity is a unified concept, yet many different types of evidence may contribute to the demonstration of validity. Validity is not based solely on any single study or type of evidence. The type of evidence on which reliance is placed will vary with the purpose of the test. The level of evidence required may vary with the potential consequences of the decisions made on the basis of the test's results. The validity evidence should be presented in a coherent validity argument supporting the inferences and actions made on the basis of the scores.

Responsibility for validity is shared by ETS, by its clients, and by the people who use the scores or other test results. In some instances, a client may refuse to supply data that is necessary for certain validity studies. ETS cannot force a client to provide data that it controls, but ETS may wish to consider whether or not to continue to provide services to the client if ETS is unable to produce evidence of validity at least to the extent required by the following standards.

Users are responsible for evaluating the validity of scores or other test results used for purposes other than those specifically stated by ETS, or when local validation is required.

Because validity is such an inclusive concept, readers of this chapter should also see, in particular, the chapters on Fairness, Test Design and Development, Reliability, Scoring, and Test Use.

Standards

Standard 4.1: Describing Test Purpose and Population

Clearly describe the construct (knowledge, skills, or other attributes) to be measured, the purpose of each test, the claims to be made about test takers, the intended interpretation of the scores or other test results, and the intended test-taking population. Make the information available to the public upon request.

Validation efforts focus on an interpretation of the test results of some population of test takers 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 test, the claims to be made about test takers, the intended interpretations of the scores or other results, and the population for which the test

is designed. For some tests, 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 tests, 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. Do not use test titles that imply that the test measures something other than what the test actually measures.

Standard 4.2: Providing Rationale for Choice of Evidence

Provide a rationale for the types and amounts of evidence collected to support the validity of the inferences to be made and actions to be taken on the basis of the test scores. For a new test, provide a validation plan indicating the types of evidence to be collected and the rationale for the use of the test.

There should be a rationally planned collection of evidence relevant to the intended purpose of the test to support a validity argument. The validity argument should be a coherent and inclusive collection of evidence concerning the appropriateness of the inferences to be made and actions to be taken on the basis of the test results.

If specific outcomes of test use are stated or strongly implied by the test title, in marketing materials, or in other test documentation, include evidence that those outcomes will occur. If a major line of validity evidence that might normally be expected given the purpose of the test is excluded, set forth the reasons for doing so.

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

- intended inferences and actions based on the test results;
- intended outcomes of using the test;
- harmful actions that may result from an incorrect inference;
- probability that any incorrect inferences will be corrected before any harm is done;
- research available on similar tests, used for similar purposes, in similar situations; and
- availability of sufficient samples of test takers, technical feasibility of collecting data, and availability of appropriate criteria for criterion-based studies.

The validity plan should be available before the first operational use of the test. Programs should monitor and document the progress made on following the validity plan.

Standard 4.3: Obtaining and Documenting the Evidence

Obtain and document the conceptual, empirical, and/or theoretical evidence that the test will meet its intended purposes and support the intended interpretations of test results for the intended populations. Compile the evidence into a coherent and comprehensive validity argument supporting the appropriateness of the inferences to be made and actions to be taken on the basis of the test results.

The evidence should, as a whole, be sufficient to indicate that the test can support the intended interpretations of test results for the intended populations and meet its intended purposes. Programs should investigate any clearly credible alternative explanations of test performance that might

undermine the validity of the inferences based on the test results, such as excessively difficult language on a mathematics test. 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.

The validity argument should present the evidence required to make a coherent and persuasive case for the use of the test for its intended purpose with the intended population. The validity argument should not be simply a compilation of whatever evidence happens to be available, regardless of its relevance. Not every type of evidence is relevant for every test. *If it is relevant to the validity argument for the test, and if it is feasible to obtain the data,* provide information in the validity argument concerning the

- procedures and criteria used to determine test content, and the relationship of test content to the intended construct;
- cognitive processes employed by test takers;
- extent to which the judgments of raters are consistent with the intended construct;
- qualifications of subject-matter experts, job incumbents, item writers, reviewers, and other individuals involved in any aspect of test development or validation;
- procedures used in any data-gathering effort, representativeness of samples of test takers on which analyses are based, the conditions under which data were collected, the results of the data gathering (including results for studied subgroups of the population), any corrections or adjustments (e.g., for unreliability of the criterion) made to the reported statistics, and the precision of the reported statistics;
- training and monitoring of raters, and/or the scoring principles 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 test, and among reported scores or other test 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 test results, subscores, and external variables (e.g., criterion variables), including the rationales for selecting the external variables, their properties, and the relationships among them;
- evidence that scores converge with other measures of the same construct and diverge from measures of different constructs;
- evidence that the test results are useful for guidance or placement decisions;
- information about levels of criterion performance associated with given levels of test performance, if the test is used to predict adequate/inadequate criterion performance;
- utility of the test results in making decisions about the allocation of resources;
- characteristics and relevance of any meta-analytic or validity generalization evidence used in the validity argument; and
- evidence that the program's claims about the direct and indirect benefits of test use are supported, including claims suggested by the title of the test.

Standard 4.4: Warning of Likely Misuses

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

No program can anticipate or warn against all the unsupported uses or interpretations that might be made of test results. However, experience may show that certain unsupported uses of the test results are likely. Programs should inform users of appropriate uses of the test, and warn users to avoid likely unsupported uses.

Standard 4.5: Investigating Negative Consequences

If the intended use of a test has unintended, negative consequences, review the validity evidence to determine whether or not the negative consequences arise from construct-irrelevant sources of variance. If they do, revise the test to reduce, to the extent possible, the construct-irrelevant variance.

Appropriately used, valid scores or other test results may have unintended negative consequences. Unintended negative consequences do not necessarily invalidate the use of a test. It is necessary, however, to investigate whether the unintended consequences may be linked to construct-irrelevant factors or to construct underrepresentation. If so, take corrective actions. Take action where appropriate to reduce unintended negative consequences, regardless of their cause, if it is feasible to do so without reducing validity.

Standard 4.6: Reevaluating Validity

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

Relevant factors include, for example, substantive changes in the technology used to administer or score the test, the intended purpose, the intended interpretation of test results, the test content, or the population of test takers.

There is no set time limit within which programs should reassess the validity evidence. A test of Latin is likely to remain valid far longer than a test of biology will, for example. The program should determine the appropriate interval between periodic reviews and provide a rationale for the time selected.

Standard 4.7: Helping Users to Develop Local Evidence

Provide advice to users of scores or other test results as appropriate to help them gather and interpret their own validity evidence.

Advise users that they are responsible for validating the interpretations of test results if the tests are used for purposes other than those explicitly stated by ETS, or if local validation evidence is necessary. Upon request, assist users in planning, conducting, and interpreting the results of local validity studies.

CHAPTER 5

Fairness

Purpose

The purpose of this chapter is to help ensure that ETS will take into account the diversity of the populations served as it designs, develops, and administers products and services. ETS will treat people comparably and fairly regardless of differences in characteristics that are not relevant to the intended use of the product or service.[1]

ETS is responsible for the fairness of the products or services it develops and for providing evidence of their fairness. There are many definitions of fairness in the professional literature, some of which contradict others. The most useful definition of fairness for test developers is the extent to which the inferences made on the basis of test scores are valid for different groups of test takers.

The best way to approach the ideal of fairness to all test takers is to make the influence of construct-irrelevant score variance as small as possible. It is not feasible for programs to investigate fairness separately for all of the possible groups in the population of test takers. Programs should, however, investigate fairness for those groups that experience or research has indicated are likely to be adversely affected by construct-irrelevant influences on their test performance. Often the groups are those which have been discriminated against on the basis of such factors as ethnicity, disability status, gender, native language, or race. (In this chapter, the groups are called the "studied" groups.) If the studied groups are too small to support traditional types of analyses, explore feasible alternative means of evaluating fairness for them.

Fair treatment in testing is addressed in laws that can change over time. Consult the Office of the ETS General Counsel periodically for the latest information about laws that may be relevant to ETS products and services. Because fairness and validity are so closely intertwined, readers of the Fairness chapter should also pay particular attention to the Validity chapter.

Standards

Standard 5.1: Addressing Fairness

Design, develop, administer, and score tests so that they measure the intended construct and minimize the effects of construct-irrelevant characteristics of test takers. For a new or significantly revised product or service, provide a plan for addressing fairness in the design, development, administration, and use of the product or service. For an ongoing program, document what has been done to address fairness in the past as well as documenting any future fairness plans.

[1] Achieving comparability may require appropriate accommodations or modifications for people with disabilities and people with limited English proficiency.

All test takers should be treated comparably in the test administration and scoring process. In either the documentation of the fairness of existing program practices or the fairness plan (whichever is appropriate) demonstrate that reasonably anticipated potential areas of unfairness were or will be addressed. When developing fairness plans, consult with clients as appropriate. 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, but differences large enough to have practical consequences should be investigated to be sure the differences are not caused by construct-irrelevant factors.

The topics to include in documentation of the program's fairness practices or the fairness plan will depend on the nature of the product or service. *If it is relevant to the product or service, and if it is feasible to obtain the data*, include information about the

- selection of groups and variables to be studied;
- reviews designed to ensure fairness, including information about the qualifications of the reviewers;
- appropriateness of materials for people in studied groups;
- affordability of the product or service;
- evaluation of the linguistic or reading demands to verify that they are no greater than necessary to achieve the purpose of the test or other materials; and
- accessibility of the product or service, and accommodations or modifications for people with disabilities or limited English proficiency.

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

- performance of studied groups, including evidence of comparability of measured constructs;
- unintended negative consequences of test use for studied groups;
- differences in prediction of criteria 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 testing for studied groups;
- evaluation of scoring procedures including the scoring of constructed responses;
- group differences in speededness, use of test-taking strategies, or availability of coaching;
- effects of different levels of experience with different modes of test administration; and
- proper use and interpretation of the results for the studied population groups.

Standard 5.2: Reviewing and Evaluating Fairness

Obtain and document judgmental and, if feasible, empirical evaluations of fairness of the product or service for studied groups. As appropriate, represent various groups in test materials. Follow guidelines[2] designed to eliminate symbols, language, and content that are generally regarded as sexist, racist, or offensive, except when necessary to meet the purpose of the product, or service.

Review materials, including tests, written products, web pages, and videos, to verify that they meet the fairness review guidelines in operation at ETS. Document the qualifications of the reviewers as well as the evidence they provide.

For tests, when sample sizes are sufficient and the information is relevant, obtain and use empirical data relating to fairness, such as the results of studies of Differential Item Functioning (DIF). Generally, if sample sizes are sufficient, most programs designed primarily for test takers in the United States should investigate DIF 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. When sample sizes are sufficient, and the information is relevant, investigate DIF for test takers with specific disabilities, and those who are English-language learners. Programs designed for nonnative speakers of English may investigate DIF for relevant subgroups based on native language. If sufficient data are unavailable for some studied groups, provide a plan for obtaining the data over time, if feasible.

Standard 5.3: Providing Fair Access

Provide impartial access to products and services. For tests, provide impartial registration, administration, and reporting of test results.

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

Standard 5.4: Choosing Measures

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

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

Standard 5.5: Providing Accommodations and Modifications

Provide appropriate accommodations or modifications for people with disabilities, and for nonnative speakers of English, in accordance with applicable laws, ETS policies, and client policies.

Tests and test delivery and response modes should be accessible to as many test takers as feasible. It will, however, sometimes be necessary to make accommodations or modifications to increase the

[2] The fairness review guidelines in operation at ETS are the *ETS Guidelines for Fairness Review of Assessments*. The *Guidelines* may be downloaded for free from **http://www.ets.org**.

accessibility of the test for some test takers. If relevant to the testing program, tell test takers how to request and document the need for the accommodation or modification. Provide the necessary accommodations or modifications at no additional cost to the test taker.

The accommodations or modifications should be designed to ensure, to the extent possible, that the test measures the intended construct rather than irrelevant sources of variation. If feasible, and if sufficient sample sizes are available, use empirical information to help determine the accommodation or modification to be made.

Accommodations or modifications should be based on knowledge of the effects of disabilities and limited English proficiency on performance as well as on good testing practices. If the program rather than the client is making decisions about accommodations or modifications, use the ETS Office of Disability Policy and the ETS Office of the General Counsel to determine which test takers are eligible for accommodations or modifications, and what accommodations or modifications they require.

If feasible and appropriate, and if sufficient sample sizes are available, evaluate the use of the product or service for people with specific disabilities and for nonnative speakers of English.

Standard 5.6: Reporting Aggregate Scores

If aggregate scores are reported separately for studied groups, evaluate the comparability of the scores of the studied groups to the scores of the full population of test takers.

If this evidence indicates that there are differences across demographic groups in the meaning of scores, examine the validity of the interpretations of the scores and provide cautionary statements about the scores, if it is necessary and legally permitted or required to do so.

Standard 5.7: Addressing the Needs of Nonnative Speakers of English

In the development and use of products or services, consider the needs of nonnative speakers of English that may arise from nonrelevant language and related cultural differences. For tests, reduce threats to validity that may arise from language and related cultural differences.

Knowledge of English is part of the construct of many tests, even if the tests are focused on another topic. For example, scoring above a certain level on an Advanced Placement® Chemistry test indicates the test taker is ready for an advanced chemistry course in an institution in which the language of instruction is English. The SAT® is designed primarily to predict success in colleges in which the language of instruction is English. For each test, indicate whether or not proficiency in English is part of the intended construct and, if so, what skills in English (e.g., reading, listening, writing, speaking, knowledge of technical vocabulary) are included.

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 and adaptation is made, describe the process and evaluate the outcome and its comparability to the original version.
- If linguistic changes are made in a test form for use by nonnative speakers of English, describe the changes in a document available to the public.

- If a test is available in more than one language, and the different versions measure the same construct, administer the test in the individual's preferred language, if that is one of the available options.
- When sufficient relevant data are available, provide information on the validity and interpretation of test results for linguistically diverse groups.
- If ETS provides an interpreter, the interpreter should be fluent in the source and target languages, be experienced in translating, and have basic knowledge of the relevant product or service.

CHAPTER 6

Reliability

Purpose

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

Reliability refers to the extent to which scores (or other reported results) on a test are consistent across — and can be generalized to — other forms of the test and, in some cases, other occasions of testing and other raters of the responses.

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

Readers of the chapter "Reliability" should also pay particular attention to the chapter "Test Design and Development."

Standards

Standard 6.1: Providing Sufficient Reliability

Any reported scores, including subscores or other reported test results, should be sufficiently reliable to support their intended interpretations.

The level of reliability required for a test is a matter of professional judgment taking into account the intended use of the scores and the consequences of a wrong decision.

Standard 6.2: Using Appropriate Methods

Estimate reliability using methods that are appropriate for the test and the intended uses of the results. Determine the sources of variation over which the test results are intended to be consistent, and use reliability estimation methods that take these sources into account.

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

- for a constructed-response or performance test, calculate statistics describing the reliability of the scoring process and statistics describing the reliability of the entire measurement process (including the selection of the tasks or items presented to the test taker and the scorers of the test taker's responses);

- for an adaptive test, provide estimates of reliability that take into account the effects of possible differences in the selection of items presented. Estimates based on resampling studies that simulate the adaptive testing process are acceptable;
- for a test measuring several different knowledge areas, skills, or abilities, use reliability estimation methods that allow for the possibility that test takers' abilities in these areas may differ;
- for a test using matrix sampling, take the sampling scheme into account;
- for a test used to classify test takers into categories (e.g., pass/fail, basic/proficient/advanced) on the basis of their scores, compute statistics indicating the form-to-form consistency of those classifications; and
- for all tests, estimate reliability statistics that are appropriate for the level of aggregation at which test results are reported (e.g., the individual student, the classroom, the school, etc.).

Reliability can refer to consistency over different sources of variation: form-to-form differences, rater differences, or differences in performance over time. Consistency over one source of variation (such as agreement between raters of the same task) does not imply consistency over other sources of variation (such as test taker consistency from task to task).

The reliability of the scores on a test depends on the sources of variability that are taken into account and the group of test takers whose scores are being considered. The reliability of decisions based on the scores depends on the part of the score scale at which those decisions are being made.

Several different types of statistical evidence of reliability can be provided, including reliability or generalizability coefficients, information functions, standard errors of measurement, conditional standard errors of measurement, and indices of decision consistency. The types of evidence provided should be appropriate for the intended score use, the population, and the psychometric models used. Estimates of reliability derived using different procedures, referring to different populations, or taking different sources of variation into account cannot be considered equivalent.

Standard 6.3: Providing Information

Provide information that will allow users of test results to judge whether reported test results (including subscores) are sufficiently reliable to support their intended interpretations. If the scoring process includes the judgment of raters, provide appropriate evidence of consistency across raters and across tasks. If users are to make decisions based on the differences between scores, subscores, or other test results, provide information on the consistency of those differences. If cut scores are used, provide information about the consistency of measurement near the cut scores and/or the consistency of decisions based on the cut scores.

Inform score users about the consistency of scores (or other test results) over sources of variation considered significant for interpretation of those results, such as form-to-form differences in content or differences between raters.

Provide score users with information that will enable them to evaluate the reliability of the test results they are using. Report reliability statistics in language appropriate for the intended audience.

Technical publications should include standard errors expressed in reported score units. When computing statistics to indicate the variability in an individual test taker's score, use procedures that take into account systematic differences between different parts of the score range (e.g., the conditional standard error of measurement). If adjusted statistics are reported, include the original values or cite available references where the original values can be found.

Standard 6.4: Documenting Analyses

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 test, and if it is feasible to obtain the data, provide information concerning

- the statistics used to assess the reliability of the scores or of other test results (e.g., reliability or generalizability coefficients, information functions, overall and conditional standard errors of measurement, indices of decision consistency, and possibly other statistics);
- the sources of variation taken into account by each statistic and the rationale for including those sources and excluding others;
- the methods used to estimate each statistic, including formulas and references for those methods;
- the population for which the reliability statistics are estimated, including relevant demographic variables and summary score statistics. (Reliability statistics may be reported separately for more than one population, e.g., students in different grades taking the same test.);
- the value of each reliability statistic in the test-taker group observed, if these values are different from the estimates for the population;
- any procedures used for scoring of constructed-response or performance tests, and the level of agreement between independent scorings of the same responses;
- any procedures used for automated scoring, including the source of responses used to calibrate the scoring engine; and
- any other pertinent aspect of the testing situation (e.g., response modes that may be unfamiliar to test takers).

Standard 6.5: Performing Separate Analyses

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

If tests are administered in long and short versions, estimate the reliability separately for each version, using data from test takers who took that version. When feasible, conduct separate reliability analyses for test takers tested with accommodations or modifications in administration or scoring.

Standard 6.6: Computing Reliability for Subgroups

Compute reliability statistics separately for subgroups of test takers when theory, experience, or research indicates there is a reason to do so.

If the same test is used with different populations of test takers (e.g., students in different grades), compute reliability statistics separately for each population.

CHAPTER 7

Test Design and Development

Purpose

The purpose of this chapter is to help ensure that tests will be constructed using planned, documented processes that incorporate advice from people with diverse, relevant points of view. Follow procedures designed to result in tests that are able to support fair, accessible, reliable, and valid score interpretations for their intended purpose, with the intended population.

Developers should work from detailed specifications, obtain reviews of their work, use empirical information about item and test quality when it can be obtained, and evaluate the resulting tests.

Meeting these standards will require test developers to work closely with others including psychometricians, scoring services, program administrators, clients, and external subject-matter experts. Because of the wide-ranging nature of their work, test developers should be familiar with all of the chapters in the *ETS Standards*, with particular emphasis on the chapters "Validity," "Fairness," "Reliability," "Scoring," and "Reporting Test Results," in addition to "Test Design and Development."

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

Standards

Standard 7.1: Describing Purpose, Population, and Construct

Obtain or develop documentation concerning the intended purposes of the test, the populations to be served, and the constructs to be measured.

Developers should know what the test is intended to measure, the characteristics of the intended test takers, and how the test is intended to be used. For some programs, the information about the intended purposes, populations, and constructs has been collected and need not be recreated. For other programs, obtaining the information may be part of the developers' task. If the information has to be obtained, work collaboratively with clients, subject-matter experts, and others as appropriate.

Standard 7.2: Providing Test Documentation

Document the desired attributes of the test in detailed specifications and other test documentation. Document the rationales for major decisions about the test, and document the process used to develop the test. Document the qualifications of the ETS staff and external subject-matter experts involved in developing or reviewing the test.

Test developers need detailed blueprints for constructing tests, and the people who evaluate and use tests need information to support decisions about test quality. Include, *as appropriate for each test,* information about the

- content, knowledge, skills, or other attributes to be measured;
- information used by test developers to determine the content, knowledge, skills, or other attributes to be measured (e.g., job analyses, curriculum surveys, course requirements, content standards, or objectives);
- extent to which the test represents the domain of content and skill, or other attributes it is designed to measure;
- test length, item formats, ordering of items and sections, and timing;
- procedures for item development, item review, pretesting, and form assembly;
- desired psychometric properties of items and the test as a whole;
- intended test administration procedures;
- rationale for the weights, if any, applied to items to obtain scores, and the rules for combining scores to obtain composite scores, if any;
- permissible variations in test administrations (e.g., paper- and computer-based, alternative formats for people with disabilities) including a rationale for the different conditions and the requirements for allowing the differences;
- results of cross-validation studies, if items are selected on the basis of empirical relationships with a criterion;
- hardware and software requirements, if the test is administered on a digital device;
- rules for item selection, starting and stopping points, scoring, and controlling item exposure, including the rationales and supporting evidence, if the test is adaptive;
- requirements for local scoring, if any; and
- procedures for score interpretation.

If separate versions of the test are sufficiently different to warrant separate documentation, such documentation should be provided.

Standard 7.3: Writing Items and Assembling Tests

Write items that meet specifications and minimize construct-irrelevant variance. Use assembly procedures or write automated assembly rules that result in tests that meet specifications and minimize construct-irrelevant variance. Follow procedures designed to ensure, to the extent possible, that the intended interpretations of test results are supported for various groups within the intended population of test takers. Evaluate the extent to which the test represents the defined construct and excludes sources of variance unrelated to the construct. If construct-irrelevant variance is found, it should be removed to the extent possible.

If the specifications are appropriate, tests that meet specifications and minimize construct-irrelevant score variance will be reliable and fair, and will lead to valid score interpretations for the intended purpose of the test with the intended population of test takers.

Standard 7.4: Obtaining Reviews

Obtain internal and/or external reviews of the test specifications, the items, the assembled test, and related materials by qualified reviewers. Document the purpose of the review, the process of reviewing, and the results of the review. To the extent that they can be ascertained, document the training, qualifications, and other relevant attributes of the reviewers.

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

- content and statistical specifications and their links to the intended interpretations of test results;
- items, including appropriateness for various groups in the intended population;
- links of items to specifications and/or to occupational tasks;
- assembled tests, or assembly rules and samples of assembled tests;
- directions for test takers; and
- ancillary materials, such as descriptive booklets and test preparation materials.

For formative tests, obtain reviews of the extent to which the test is likely to provide useful information to aid learning.

Obtain substantive contributions from qualified persons who represent relevant perspectives, professional specialties, population groups, and users of the results of the test. For tests 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, tests, directions, and ancillary materials should be performed by people who are familiar with the specifications and purpose of the test, the subject-matter of the test as necessary, and the characteristics of the test's intended population.

Important aspects of the review include

- content accuracy;
- suitability of language;
- match of items or tasks to specifications;
- accessibility and fairness for population groups;
- editorial considerations;
- completeness and clarity of directions and sample items;
- completeness and appropriateness 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 interpretations made of the test scores.

Standard 7.5: Pretesting

When feasible, pretest items with test takers that represent, to the extent possible, the intended population for the test. Document the sampling process and the characteristics of the resulting sample. Document the statistical procedures used to evaluate the items and the results of the analyses, including, as appropriate, the fit of the model to the data. If pretesting is not practicable, use small-scale pilot tests, and/or collateral information about the items, and/or a preliminary item analysis after an operational administration but before scores or other test results are reported.

Pretesting is a means of obtaining information about the statistical characteristics of new items, and possibly the adequacy of testing procedures. Sometimes pretesting is used to obtain data for predicting final form statistical characteristics. Terminology is not well standardized. Generally *pretest* and *field test* are used synonymously, and *pilot test* refers to the administration of the items to small samples of test takers. If items are substantially revised based on pretest results, it is good practice to pretest the revised items, when it is feasible to do so.

If pretesting is not feasible, the collateral information that may be used includes, for example, the number of operations and level of cognition required to respond to an item, closeness of the distracters to the key, and performance of similar items used with similar populations. 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 groups within the population of test takers.

Standard 7.6: Evaluating Operational Tests

Evaluate the performance of tests after the first operational administration at which sample sizes are sufficient.

Carry out timely and appropriate analyses, including analyses for reliability, intercorrelation of sections or parts, and indications of speededness. Evaluations of adaptive tests may be based on sample forms after simulated administrations, but data on real test takers should be evaluated when possible. Evaluations of some tests may require the collection of data over time, as the tests are used.

Evaluate representative test editions in terms of the degree to which they meet their psychometric specifications. 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: Obtaining Periodic Reviews

Periodically review test specifications, active items, tests, and ancillary materials to verify that they continue to be appropriate and in compliance with current applicable guidelines. Revise materials as indicated by the reviews. Notify test takers and test users of changes that affect them.

Programs should determine, in collaboration with clients, as appropriate, review periods, which will depend on the nature of the test, and provide a rationale for the selected interval.

For adaptive tests, periodically evaluate, as appropriate, sample test forms, the adequacy of the fit of item-response models, the size and security of item pools, and the adequacy of the computerized procedures 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: Contributing to the Validity Argument

Collaborate with researchers and psychometricians in collecting evidence of validity for the validity argument.

Much of the evidence in the validity argument is the responsibility of test developers and is documented as test developers comply with the standards relevant to their work. As the evidence is collected, augment the validity argument. Work with psychometricians and researchers to help ensure that each of the claims made about test takers on the basis of the scores is supported by a cohesive, comprehensive, and convincing validity argument. (Please see Chapter 4, "Validity," for more information about the validity argument.)

CHAPTER 8

Equating, Linking, Norming, and Cut Scores

Purpose

The purpose of this chapter is to help ensure that test scores meant to be comparable will be correctly adjusted to make them comparable, that normative data will be meaningful, and that cut score studies designed or implemented by ETS will use rational, clearly described procedures.

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

Standards

Standard 8.1: Using Appropriate Equating or Linking Methodologies

When scores are meant to be comparable, use appropriate methods for equating scores on alternate forms of a test and for linking scores on different tests.

There are different ways to link scores. Different types of equating or linking are appropriate for different circumstances. Strict equating models are appropriate for scores on alternate forms of the same test that are deemed interchangeable in terms of content and statistical characteristics. Statistical linking methods are appropriate for comparing scores on different tests in the same test-taker population. Judgmental methods are appropriate for comparisons of scores on tests intended for different populations (e.g., at different grades or measuring different constructs).

Standard 8.2: Documenting Equating or Linking — Population and Comparability

Specify the test-taker population in which the scores connected by the equating or linking study are to be interpreted as comparable, and state the definition of comparability.

Probably the most common definition of comparability is that scores on two tests are comparable in a group of test takers if the scores on both tests represent the same relative position in that group of test takers. If this is the relevant definition of comparability, specify the group (e.g., all test takers taking the test under standard conditions) and the way in which relative position is determined (e.g., by percentile rank). Equating or linking procedures based on Item Response Theory (IRT) often define scores on two tests as comparable if they are the expected scores for test takers of the same ability in some population of potential test takers. Describe the characteristics of the test-taker samples included in the equating or linking study and the similarity of the test forms being equated or linked.

Standard 8.3: Documenting Equating or Linking — Data Collection Design

Describe the data collection design for the equating or linking study and state explicitly the assumptions implied by the use of that design.

If the study relies on the equivalence of groups taking different forms of the test, explain how the groups were selected or otherwise determined to be of equal ability. If the study relies on anchor items, describe the content and the statistical characteristics of the set of anchor items and the ways (if any) in which it differs from the set of items on the test.

Also determine the statistical relationships between scores on the anchor and scores on each form of the test. If equating or linking is done by IRT methods, describe the sampling, calibration, and equating or linking processes used to develop and maintain the IRT scale.

Standard 8.4: Documenting Equating or Linking — Statistical Procedures

Specify the statistical procedures used to determine the equating or linking conversion and the assumptions underlying those procedures.

Describe the procedures in sufficient detail to allow knowledgeable people to evaluate and replicate the studies. The description need not include all the details, if it includes references to publicly available sources where the necessary details can be found.

Standard 8.5: Documenting Equating or Linking — Results

Document the results of the equating or linking study.

Include estimated standard errors of equating or other statistics indicating the sample dependence of the results of the equating or linking study. If the equating or linking is based on IRT, provide information about the adequacy of fit of the model to the data.

Standard 8.6: Equating Test Scores

In documenting the equating of scores intended to be used interchangeably, provide a clear rationale for the interchangeable use of the scores, a precise description of the method by which the equating functions were established and, if possible, information on the accuracy of the equating functions.

Show that forms to be equated are measuring the same construct and document the content and statistical similarities of the forms.

Standard 8.7: Using Norm Groups

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

Documentation of norming studies should include the dates of the studies, definition of the populations sampled, the procedure used to draw the samples, sample sizes, participation rates, and any weighting or smoothing procedure used to make the sample data better represent the population.

If the norms information includes "program norms" based on the test takers for whom data are available, inform score users of the limitations of those norms.

If published norms are likely to differ substantially from local norms, 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.

Standard 8.8: Designing Cut Score Studies

If ETS is involved in designing a cut score study, provide users with the information they need to choose an appropriate methodology and study design.

Generally, cut scores should be set by authorities who have the responsibility to do so under the laws of some jurisdiction. ETS designs and implements standard-setting studies at a client's request, and provides other data, such as score distributions, to assist the client in setting a cut score.

Design the cut score study so that any judgmental process involved calls for judgments that reflect the raters' knowledge and experience. Provide relevant empirical information (e.g., item difficulty, relationships between test performance and relevant criteria), if the information is available.

Standard 8.9: Implementing Cut Score Studies

When implementing a cut score study, choose appropriate samples of raters from relevant populations, and train the raters in the method they will use.

Make sure that the raters in a cut score study understand the purpose of the test 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. Describe any study features used to improve the quality of the ratings, such as multiple rounds of ratings, provision of interim results, rater discussions, frequent rater feedback questionnaires, and provision of item difficulty data. If the data are available, provide the raters with estimates of the effects of setting the standard at various points.

Standard 8.10: Documenting Cut Score Studies

Document the cut score study in sufficient detail to allow knowledgeable people to evaluate and replicate the study. Summarize the logical and/or empirical evidence supporting the classifications made on the basis of the cut scores.

The documentation of the cut score study should include information about criteria for the selection of raters, and how they were trained. Describe any data provided to the raters and any ways in which raters were allowed to confer with each other while making their judgments. Describe the procedure for combining the raters' individual judgments. Include full descriptions of the procedures followed and the results. When feasible, provide estimates of the variation that might be expected in the cut scores if the study were replicated with different raters.

The "validity argument" described in Chapter 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 scores.

CHAPTER 9

Test Administration

Purpose

The purpose of this chapter is to help ensure that tests will be administered in an appropriate manner that provides accurate, comparable, and fair measurement.

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

The shift of test administration to a variety of digital devices presents new challenges in maintaining standardization. The standards in this chapter apply regardless of the apparatus used to administer the test.

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

Standards

Standard 9.1: Providing Information for Test Administrators

Provide those who administer tests 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 tests information about the

- purpose of the test, the population to be tested, and the information needed to respond to expected questions from test takers;
- reasons why it is important to follow test administrations directions carefully;
- qualifications required to administer the tests;
- required identifying information for test takers, admittance procedures, timing, and directions;
- variety of devices on which the test may be administered;
- devices, materials, aids, personal possessions or tools that are required, permitted, or prohibited;
- maintenance of appropriate security procedures;
- actions to take if irregularities are observed;
- methods of communicating with ETS to report problems;
- operation of equipment and software as needed for the administration;

- provision of approved accommodations or modifications; and
- resolution of problems that may delay or disrupt testing.

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

Standard 9.2: Providing Information for Test Takers

Tell test takers what they can expect in the registration processes, and during the test administration.

Provide clear information to test takers, including how to

- register for the test;
- request accommodations or modifications;
- present appropriate identification materials;
- distinguish among the materials, tools, personal possessions and aids that are required, permitted, or prohibited during the test;
- take the test and make responses;
- observe the rules dealing with proper behavior during the administration; and
- report problems or complaints about registration services, the administration, and/or the items.

Inform test takers of the actions that constitute misconduct during the test administration and warn them about the consequences of misconduct.

Standard 9.3: Providing Appropriate Testing Environment

Provide a reasonably comfortable testing environment in locations reasonably accessible to the intended populations.[3] Follow established procedures for test administration.

To reduce the introduction of construct-irrelevant variance during test administrations, administer tests in an environment with appropriate temperature control, reasonable furniture, adequate lighting and work space, low noise levels, well-functioning equipment, and as few disruptions as reasonably possible. As necessary, provide access to a testing site for people with disabilities. Programs should monitor samples of administration sites, as appropriate and feasible, to verify that tests are being administered as specified.

Standard 9.4: Protecting Secure Tests

For secure tests provide clear directions on the procedures necessary to maintain security before, during, and after the administration process.

The level of security needed to protect the integrity of the test and scores, or other test results, depends on the purpose and nature of the test. Security may not be a concern for some tests. Access to digital devices containing secure test material or to secure paper test material should be restricted

[3] This does not apply if tests are administered outside of an official test center, such as in the test taker's classroom or home.

to authorized people before, during, and — unless the material is to be disclosed following the test administration — after test administration.

Standard 9.5: Maintaining Score Integrity

To the extent feasible, eliminate opportunities for test takers to attain scores by fraudulent means.

Work with clients, as appropriate, to establish procedures to protect the integrity of scores. The procedures should be designed to eliminate misconduct such as impersonation, obtaining answers from others, obtaining prior knowledge of secure test items, using unauthorized aids, obtaining unauthorized assistance, obtaining copies or images of secure test items, and invading computer systems containing test materials or answer keys. Tell people who administer tests how to monitor test takers and take appropriate action if misconduct or irregularities are observed.

Standard 9.6: Maintaining Comparability

If a test is administered on different digital devices, ensure to the extent possible that the scores obtained on different devices are comparable.

Perform comparability studies to help ensure that test takers are not disadvantaged by use of any of the allowable devices for test administration.

CHAPTER 10

Scoring

Purpose

The purpose of this chapter is to help ensure that ETS will establish, document, and follow procedures that will result in accurate and consistent scoring of test takers' responses to the tasks on ETS tests.

Scoring of multiple-choice answer sheets, constructed responses, or complex performances, scored by human raters or by machines, should follow documented procedures and should be checked for accuracy.

Standards

Standard 10.1: Developing Procedures for Human Scoring

If the scoring involves human judgment, develop clear, complete, and understandable procedures and criteria for scoring, and train the raters to apply them consistently and correctly.

Raters need training to apply the scoring rubric consistently and correctly. Rater training should include "benchmarks" (responses typical of different levels of performance at each score point). The training should include enough examples to make the intended scoring standards clear to the raters. Expert raters should be available to assist when a rater has difficulty rating a response.

Standard 10.2: Monitoring Accuracy and Reliability of Scoring

Monitor and document the accuracy and reliability of scoring, and correct sources of scoring errors.

If the rating process requires judgment, the accuracy of ratings can best be judged by comparing them with ratings assigned to the same responses by expert raters. Scoring by human raters can be monitored by "back-rating" — periodically sampling responses to be rescored by an expert rater. It can also be monitored by sampling responses, having them scored in advance by expert raters, and inserting these "validity" responses into each rater's workflow. Machine scoring of selected-response answer sheets can be monitored by hand-scoring a sample of the answer sheets.

If scoring is done by IRT methods, describe the sampling, calibration, and linking processes used to develop and maintain the IRT scale. If local scoring is intended, ancillary test materials should contain information regarding the procedures to use to monitor the accuracy and reliability of scoring.

Standard 10.3: Using Automated Scoring

If the test includes automated scoring of complex responses, use human raters as a check on the automated scoring.

The extent to which human raters are required will vary with the quality of the automated scoring and the consequences of the decisions made on the basis of the scores. If the automated scoring has been shown to be an acceptable substitute for the human scoring, or if the test results will not be used to make decisions with important consequences, or if the automated score contributes only a small portion of the total score, it will be sufficient to limit the human scoring to a sample of the responses. If, however, the automated scoring has not been shown to be an acceptable substitute for human scoring, and if the score will be used to make decisions with important consequences, then use a human rater as a check on every automated score. Any disagreements between the human and automated scorings should be resolved by another human rater.

Standard 10.4: Documenting Scoring Procedures

Document the scoring procedures.

Document the procedures by which each test taker's performance is evaluated and translated into a reported score. If the scoring includes ratings by human raters, document the procedures used to train the raters, to assign responses to raters for scoring, and to monitor the scoring process. If the scoring includes automated scoring of constructed responses, document the process of calibrating the scoring engine, including the selection and scoring of the responses used in the calibration process. If the reported score is a weighted composite of several scores, describe the weighting process and the rationale for its use.

CHAPTER 11

Reporting Test Results

Purpose

The purpose of this chapter is to help ensure that the scores, other test results, and interpretive information that ETS provides are understandable and meaningful to the intended recipients.

Readers of this chapter should also pay particular attention to the chapter "Equating, Linking, Norming, and Cut Scores."

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

Standards

Standard 11.1: Explaining Scores and Other Test Results

Provide test score users with the information they need to understand the meaning and the limitations of the scores and any other test results reported. Provide information to help score users and the public avoid misinterpretation of scores or other test results for individuals or for groups. Warn intended recipients of any likely misinterpretations of the reporting scale.

Test results can include scores, score profiles, diagnostic information, status labels such as *Proficient,* aggregated data, imputed scores, and summaries of individual or group performance. Recipients of this information can 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 or similar types of information at different levels of technical detail.

Programs that produce test results in which there is a legitimate public interest should provide information that will help the news media and the general public understand the results.

Information that is necessary to understand test results should be provided at the same time as the test results.

- If raw scores or percent correct scores are reported, inform intended recipients of the limitations of those scores with respect to comparability across tests and across different forms of the same test.
- If test results are reported as labels, use the least stigmatizing labels consistent with accuracy.
- If pass/fail classifications are reported, give failing test takers information about their performance relative to the cut score, if it is feasible and the information is helpful to the test takers.

- If test results for individual test takers are reported in terms of broad categories, include information that indicates the test taker's position within the category, if the information is helpful to the test takers and/or score users.
- If any particular misinterpretation of a score scale is likely, warn score users to avoid that misinterpretation.
- If there are any changes in the test that may affect the interpretation of the scores, inform score users of the changes.
- If scores for different tests are reported on the same scale but are not comparable, caution users that those scores are not comparable.

Standard 11.2: Reporting on Appropriate Scales

Report test scores on scales that are appropriate for the intended purpose of the test and that discourage misinterpretations. Provide a rationale for the choice of each score scale and document the procedures for converting raw scores to those scales.

Reporting scales differ in the number of points used and in the relationship of raw scores to scaled scores. For example, a scaled score of 150 could be the lowest reported score of one test, the mean score of a second test and the highest reported score of a third test. A change of 10 points could be an immaterial change on one score scale, but a change of several score levels on some other score scale.

Provide a statement of the reasons for choosing the score range and interval of the score scale. For established programs, the reason may simply be that maintaining the continuity of the historical scale is necessary to avoid confusing score users.

Standard 11.3: Evaluating Stability of the Reporting Scale

Check the stability of the reporting scale whenever the test or the test-taker population changes significantly. A score scale is stable if the meaning of the scores reported on that scale does not change over time. If the meaning of scores changes significantly, take steps to minimize misinterpretations. Check the stability of the reporting scale periodically if appropriate use of the scores or other test results depends on the stability of the scale.

If a change to the test or to the test-taker population changes the meaning of the test results, it is important to minimize confusion between the old and new meanings. One alternative is to change the scale used to report the results. Another alternative is to communicate the differences clearly to score recipients.

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

Standard 11.4: Providing a Frame of Reference

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

The frame of reference might include information from norms studies, carefully selected and defined program statistics, research studies, or surveys of experts. Clearly describe the nature of the groups on which the information is based. Distinguish between carefully sampled representative norm groups and nonrepresentative groups such as self-selected groups ("program norms") or convenience samples.

Standard 11.5: Reporting Subgroup Scores

If information about scores or other test results is reported for population subgroups, report it in a way that encourages correct interpretation and use of the information.

Accompany the information with an explanation that will enable the recipients to interpret it correctly. Avoid developing separate information for groups so small as to make the data potentially misleading.

CHAPTER 12

Test Use

Purpose

The purpose of this chapter is to help ensure that ETS will describe and encourage proper test use and discourage common misuses of the test.

ETS will promote proper use of tests and help score recipients use tests fairly and appropriately, in accordance with supporting evidence. Readers of this chapter should be familiar with Chapter 4, "Validity."

Standards

Standard 12.1: Providing Information

Provide intended users of ETS tests with the information they need to evaluate the appropriateness of tests for their intended purposes. Provide the users with opportunities to consult with ETS staff about appropriate uses of tests.

Provide users of ETS tests with information such as the

- intended purpose of the tests and the intended populations of test takers;
- general content and format of the tests;
- difficulty and reliability of the tests, and validity of the test scores;
- availability and applicability of normative data;
- administration and scoring requirements;
- policies for data retention and release;
- representative research relevant to the tests; and
- methods for obtaining answers to questions about the uses of the tests.

Standard 12.2: Using Tests Properly

Encourage proper test use and appropriate interpretations of test results. Caution users to avoid common, reasonably anticipated misuses of the test.[4]

Proper test use is the responsibility of the institutions and people who employ test scores to help make decisions. ETS cannot monitor every use of its tests. However, programs should encourage proper test use by taking such actions as informing test users

- how to use scores or other test results in conjunction with other relevant information (if such information is available and useful);

[4] A misuse is an invalid use of test scores resulting in harmful effects.

- how to evaluate score differences between individuals (or between subscores for the same individual);
- of alternative, plausible explanations for poor performance by test takers;
- of the need to explain and justify the use of any accountability index based on test scores, such as a value added model, and of the need to investigate the reliability, validity, and fairness of the index;
- of the need to consider whether or not test takers have had the opportunity to learn the content and skills measured in tests used for graduation or promotion decisions;
- of the need to allow a reasonable number of opportunities to succeed for students who must pass a test to be promoted or granted a diploma, or for individuals seeking certification; and
- of the need to become familiar with the responsibilities of test users.

Discourage misuses of scores or other test results by warning users of likely misuses and how to avoid them.

Standard 12.3: Investigating Possible Misuse

Programs should investigate credible allegations of test misuse, when feasible. If test misuse is found, inform the client and the user. Inform the user of the appropriate use of the test. If the misuse continues, consult with the client as appropriate concerning suitable corrective actions.

The appropriate actions may include withholding scores or other test results from score recipients who persist in harmful misuse of an ETS test. Controversial uses of tests may result in allegations of misuse, but a controversial use of a test is not necessarily a misuse of the test.

Standard 12.4: Evaluating Test-Based Decisions

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

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

Standard 12.5: Avoiding Use of Outdated Scores

Establish a policy to deter use of outdated scores or other outdated test results. State the time during which test results will remain valid and will continue to be reported.

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

CHAPTER 13

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 endeavor to protect their rights during all phases of the testing process.

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

Standards

Standard 13.1: Describing Rights and Responsibilities

Inform test takers of their rights and responsibilities.

Statements of the rights of test takers should include, for example, such issues as

- their right to information about the nature and purpose of the test;
- any recourse they have if the test 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 test results reported by ETS, unless they have been canceled or the right has been waived.

Statements of the responsibilities of test takers should include, for example, such issues as

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

Standard 13.2: Providing Access and Information

Provide all test takers with respectful and impartial treatment, appropriate access to test services, and information about the test and the test 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. Make test preparation materials, sample tests, and information about the testing process available to test takers in appropriate formats. Test preparation materials should be accessible to the extent possible, and relevant to the tests currently being administered.

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

- the characteristics of the test and item types;
- the testing fees (if any), eligibility requirements for fee waivers, and how to request fee waivers;
- how their personal information (including scores and other results, credit card information, and other identifying information) will be protected;
- how to request appropriate accommodations or modifications;
- the intended uses of any scores or other test results, as well as the conditions under which results will be reported and to whom;
- how long scores will remain available and usable;
- the score or combinations of scores required to pass tests used with passing scores, if the information is available;
- the advantages and disadvantages of each mode of test, if multiple modes are offered;
- which materials or personal possessions are required for the test, which are permitted, and which are prohibited;
- appropriate test-taking strategies;
- how often, how soon, and under what conditions the test may be taken again;
- whether there are scoring procedures that may affect their results (e.g., if partial credit is given), 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 the test or samples of items, and whether it is possible to obtain a record of responses or have the test 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: Obtaining Informed Consent

Obtain informed consent from test takers as necessary.

Administer tests or release personally identifiable scores or other test results only when test takers (or the parents or guardians of minor test takers) 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. Inform test takers (or the parents or guardians of minor test takers) if personally identifiable information about the test takers will be used for research purposes.

Standard 13.4: Registering Complaints

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

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

Standard 13.5: Questioning Scores

If ETS cancels scores or does not report scores within the normally expected reporting time because of suspected administrative irregularities, invalidity, or misconduct, follow documented procedures approved by the ETS Office of the General Counsel designed to protect the rights and privacy of test takers whose scores are affected. Allow test takers to offer relevant evidence as appropriate, and consider such evidence in determining the validity of questioned scores.

Programs with secure tests should have documented policies, approved by the client, as appropriate, and by ETS's Office of the General Counsel, describing procedures in the case of scores questioned on the basis of administrative irregularities, invalidity, or possible misconduct. 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 scores, and any options that test takers may select for resolving the matter. When appropriate, consider reasonably available relevant information that includes material the test taker may choose to provide.

Glossary

Accommodation — A change to a test, or to its administration, to assess the intended construct for a person with a disability or for an English-language learner. An accommodation does not change the intended construct. See **Construct, Disability**. Compare **Modification**.

Adaptive Test — A test in which the items presented depend on the test taker's responses to previous items. In general, correct responses lead to more difficult items, and incorrect responses lead to less difficult items. The goal is to present items that provide the most information about the test taker's level of ability. Most adaptive tests are administered on a computer and are referred to as Computerized Adaptive Tests, or CATs.

Adjusted Coefficient — A statistic that has been revised to estimate its value for a group other than the sample on which it has been calculated or under conditions other than those in which the data were obtained.

Administration Mode — The method by which a test is presented to the test taker, including, for example, printed booklets, Braille booklets, American Sign Language, computer display terminals, audio files, and videos. See **Response Mode**.

Alternate Forms — Different editions of the same test, written to assess the same skills and types of knowledge at the same level of difficulty, but with different tasks, questions, or problems.

Alternate Forms Reliability — An estimate of reliability based on the correlation between alternate forms of a test 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.

Analytic Scoring — A procedure for scoring responses on a constructed-response test, in which the rater awards points for specific features or traits of the response. Compare **Holistic Scoring**.

Anchor Test — A short test or portion of a test 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 a test.

Assessment — See **Test**.

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 the OPSC, 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 Scoring of Constructed Responses — The use of a computerized procedure to generate a score for an essay or other constructed response. (The computer program that implements the procedure is often referred to as a "scoring engine.") See **Constructed Response**.

Bias — In general usage, unfairness. In technical usage, the tendency of an estimation procedure to produce estimates that deviate in a systematic way from the correct value. See **Fairness**.

Calibration — (1) In item response theory, calibration is the process of estimating numbers (parameters) that measure the difficulty and discrimination of each item. (2) In the scoring of constructed responses by human raters, calibration is the process of establishing consistency between raters in their standards for awarding each possible rating. (3) In the automated scoring of constructed responses, calibration is the process of adjusting the scoring engine (the computer program) to reproduce, as closely as possible, the ratings given to a set of previously scored responses (the calibration sample). See **Automated Scoring of Constructed Responses, Holistic Scoring, Item Response Theory**.

Canceled Score — A canceled score is a score that is not reported on a test that has been taken, or a score that 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, invalidity, misconduct, or other reasons. See **Irregularity**.

Certification — Official recognition by a professional organization that a person has demonstrated a high level of skill or proficiency. Sometimes, certification is used as a synonym for licensing. Compare **Licensing**.

Client — An agency, association, foundation, organization, institution, or individual, that commissions ETS to provide a product or service.

Coaching — Short-term instruction aimed directly at improving performance on a test. Coaching can focus on test-taking strategies, on knowledge of the subject tested, or on both.

Common Items — A set of test questions included 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**.

Comparable Scores — Scores that allow meaningful comparisons in some group of test takers. The term usually refers to scores on different tests or on different forms of the same test. Compare **Equating**.

Composite Score — A score that is the combination of two or more scores by some specified formula, usually a weighted sum.

Computer-Based Test — Any test administered on a computer.

Construct — The set of knowledge, skills, abilities, or traits a test 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-Irrelevant Variance — Differences between test takers' scores that are caused by factors other than differences in the knowledge, skills, abilities, or traits included in the construct the test is intended to measure. See **Construct, Variance**.

Construct Validity — All the theoretical and empirical evidence bearing on what a test is actually measuring, and on the qualities of the inferences made on the basis of the scores. Construct validity was previously associated primarily with tests of abstract attributes, such as honesty, anxiety, or need for achievement, rather than tests 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 test scores. The concept of various types of validity has been replaced

by the concept of various types of evidence supporting a unified concept of validity. See **Validity**. Compare **Content Validity, Criterion-Related Validity**.

Constructed Response — A response (to a test question, task or exercise) generated by the test taker rather than selected from a list of possible responses. Compare **Selected Response**.

Content Validity — The aspect of construct validity that emphasizes evidence bearing on the appropriateness of the knowledge, skills, and abilities measured by a test. Are the important areas of the domain (of knowledge and skills to be tested) represented by appropriate numbers of items? Are important areas of the domain excluded? Is material outside the domain included? The concept of various types of validity has been replaced by the concept of various types of evidence supporting a unified concept of validity. See **Domain, Validity**. Compare **Construct Validity**.

Criterion — That which is predicted by a test, such as college grade point average or job performance rating.

Criterion-Related Validity — The aspect of construct validity that emphasizes evidence bearing on the statistical relationships between test scores and other variables of interest. Often, the relationship is predictive. Criterion-related validity evidence 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. The concept of various types of validity has been replaced by the concept of various types of evidence supporting a unified concept of validity. See **Criterion**, **Validity**.

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 — Interactions of ETS staff with customers for the purpose of increasing customer satisfaction. See **Customer Satisfaction**.

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 **Standard**.

Differential Item Functioning (DIF) — Any tendency of a test item to be harder or easier for members of a particular group of test takers than for equally able members of another group. Generally, in measures of DIF, members of different groups are considered equally able if they receive the same score on a test. In most DIF analyses, the groups of test takers are defined on the basis of gender, race, or ethnicity. Compare **Impact**.

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

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

Documentation — Tangible evidence, generally in written form, of compliance with a standard. Most of the documentation specified in these standards is for 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 a test so that the scores may be used interchangeably. See **Anchor Test**, **Common Items**, **Conversion Parameters**. Compare **Linking**.

ETS Board of Trustees — The ETS Board of Trustees is the governing body of ETS.

Fairness — The extent to which a product or service is appropriate for members of different groups. For tests, 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 appropriate representation of different groups among those receiving benefits based on test scores, even if the groups have different scores, on average. One useful definition of fairness is that a test is fair if any group differences in performance are derived from construct-relevant sources of variance. The existence of group differences in performance does not necessarily make a test unfair, because the groups may differ on the construct being measured. See **Bias**, **Validity**.

Form — An edition of a test. See **Alternate Form**.

Formative Evaluation — Tests given during instruction to help shape ensuing instruction. Also, 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**.

Holistic Scoring — A procedure for scoring responses on a constructed-response test, in which the rater makes a single overall judgment of the response. Compare **Analytic Scoring**.

Impact — A raw difference between groups in percent correct on an item, or in scores or passing rates on a test. Compare **Differential Item Functioning (DIF)**.

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

Intended Population — The test takers for whom a test has been designed to be most appropriate.

Irregularity — Problem or disruption in connection with a test administration.

Item — A test question, problem, or task.

Item Analysis — Statistical analysis of the responses of test takers to items, done for the purpose of evaluating the items, rather than the test takers. The results typically include a measure of the difficulty of the item, the number of people choosing each of the options in a multiple-choice item, 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 (e.g., ability) and certain characteristics of the items (e.g., difficulty).

Item Type — The observable format of a question. At a very general level, "item type" may indicate only whether the item is multiple-choice or free response. Items that require different systems processing for capture of responses, such as an audio response vs. a text response or a single selection vs. multiple selections, are considered different item types. At a much finer level of distinction, "item type" may indicate, for example, whether the item calls for synonyms or antonyms; in this usage, the term overlaps with "task type."

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

Key — The correct answer to a test question, or a listing of the correct responses to a set of test questions.

Licensing — The granting by a government agency of permission to practice an occupation or profession, based on evidence that the applicant has the knowledge and skills needed to practice that occupation without endangering the public. Compare **Certification**.

Linguistic Demands — The reading and/or listening ability necessary to comprehend the questions or tasks on a test, and the writing and/or speaking ability necessary to respond.

Linking — The general term for making scores on different tests comparable in some way. 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 — Score distributions and related statistics for a single institution or a 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 Test — A test that is given by an institution at a time and place of the institution's own choosing.

Matrix Sampling — A method of test 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 test taker has to respond to all of the items that are administered.

Mean — The arithmetic average. The sum of a set of numbers divided by the number of numbers.

Measurement Error — In the context of testing, 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 a test, or whether the rater 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**.

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

Misuse — An invalid use of test scores resulting in harmful effects.

Modification — A change to a test or its administration to make the test accessible for a person with a disability or for an English-language learner. A modification changes the intended construct. See **Accommodation**, **Construct**, **Disability**.

Multiple-Choice Test — A test in which the test taker selects the correct response to an item from a limited number of answer choices.

Norm Group — A group of test takers used as a basis for comparison. The scores of individual test takers are given meaning by comparison to the distribution of scores in the norm group. See **Normative Scale**, **Norms**.

Normative Scale — A score scale that is defined in terms of the performance of a norm group. In many cases, the scale is defined by specifying values for the mean and the standard deviation of the scaled scores of the norm group. See **Norm Group**, **Norms**.

Norms — Performance data for a norm group, usually expressed as the percentile rank of each possible score on the test. Norms 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 test taker receives by taking a particular form of a test at a particular administration. Compare **True Score**.

Operational Administration — The use of a test 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 that value computed from a sample drawn from the population. (2) In item response theory, one of the characteristics of an item that determine its difficulty for a test taker with a given amount of ability.

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

Percent Correct Score — The number of questions answered correctly, divided by the number of questions on the test. Percent correct scores are not comparable across different tests. For example, 60 percent correct on a difficult test may indicate a higher level of knowledge than 80 percent correct on an easier test. Compare **Percentile Rank**.

Percentile Rank — The percent of some defined group who scored below a particular score on a test. For example, a score at the 84th percentile of a group is higher than the scores of 84 percent of the test takers in that group. (Percentile ranks are sometimes defined to include half the test takers who scored exactly at the given score, in addition to those scoring lower.) Compare **Percent Correct Score**.

Performance Test — A test 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 — Small-scale preliminary try out of new test questions, a new test form, or a new type of test. Pilot testing often involves observation and interviewing of test takers. (Sometimes used as a synonym for pretesting.) Compare **Pretest**, **Operational Administration**.

Pool — The set of items from which a test or group of tests will be assembled.

Population — All the members of some defined group, such as third-grade students in the United States. Most populations are too large for every member to be tested. Compare **Sample**.

Population Group — A part of a larger population that is defined on the basis of a characteristic such as gender, race or ethnic origin, training or formal preparation, geographic location, income level, disability, or age. Also called population subgroup.

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

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.

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

Preliminary Item Analysis — An item analysis performed after a test has been administered operationally 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 a test to gather data on item or test characteristics, such as difficulty and discrimination of items. Some authors use "field test" as a synonym for "pretest." See **Discrimination**. Compare **Operational Administration**.

Profile — A pattern of scores of the same test taker on different tests or on different subscores of the same test. 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.

Rater — A person or computerized procedure that assigns a score to a constructed response. See **Score**.

Raw Score — A test score that has not been adjusted to account for the difficulty of the items. Types of raw scores include the number of items answered correctly, the percentage of items answered correctly, the sum of ratings assigned to the test taker's responses, and weighted sums of these scores.

Registration — The process of enrolling to take an ETS test.

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

Reliability — An indicator of the extent to which scores will be consistent across — and generalize to — different administrations, and/or administration of alternate forms of the test, and/or different raters. See **Alternate Form Reliability**, **Test-Retest Reliability**, **True Score**, **Variance**.

Reliability of Classification — The extent to which test scores would be consistent in assigning a test taker to the same category (such as basic, proficient, or advanced), if the test taker were retested with the same test or an alternate form of the test, assuming no relevant change in the test taker's abilities. 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 situation in which an analysis does not include data from those people who would have had the highest or lowest scores on one (or more) of the variables. In this situation, the sample in the analysis differs systematically from the population that the results are intended to describe. See **Population**, **Sample**, **Variance**.

Rubric — A set of rules and guidelines for assigning scores to constructed-response or performance items. Generally, the rubric describes the attributes of responses associated with each score level. Often rubrics are accompanied by examples of responses at various score levels.

Sample — A subset of a larger population, selected to represent that population for statistical purposes. For example, a few hundred high schools may be selected to represent all of the 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.

Scaled Score — A test score expressed on a scale that is different from that of the raw score, usually because it has been adjusted to take into account the difficulty of the items or tasks on a specific form of the test. Compare **Raw Score**.

Scaling — The process of transforming raw scores to scaled scores. See **Raw Score**, **Scaled Score**.

Score — A number or label that represents a test taker's performance on a test.

Score Recipient — A person or institution to whom test scores are sent. Those test scores can be the scores of individual test takers or summary data for groups of test takers.

Score Scale — The set of numbers that are possible values for scores on a test.

Secure Test — A test in which the items or tasks are not to be available to test takers prior to the administration of the test to them.

Service Standards — A specified value for evaluating interactions with customers, such as the average number of rings before an 800 number is answered, or speed of order fulfillment.

Specifications — Detailed documentation of the intended characteristics of a test, including but not limited to the content and skills to be measured, the numbers and types of items, the level of difficulty and discrimination of the items, the timing, the presentation and response formats, and the layout of the test.

Speededness — The extent to which test takers' performance on a test is affected (negatively) by time limits. Some tests are speeded on purpose, if speed of response is an aspect of the construct being measured. For most tests, however, speededness is not a desirable characteristic. See **Construct**.

Standard — (1) A cut score or a required 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 that indicates the amount of variation in a set of scores, expressed in the same units as the scores themselves. The standard deviation can be interpreted as the typical distance of an individual score from the mean score. It can be computed by squaring the difference between each individual score and the mean, averaging those squared differences, and taking the square root of that average. The standard deviation is the square root of the variance. See **Variance**.

Standard Error of Estimate — A statistic that indicates the extent to which an estimate of a quantity tends to vary over samples from the same population. Specifically, it is the standard deviation of the distribution of the estimates that would be obtained from all possible samples of a given size. See **Standard Deviation**.

Standard Error of Measurement — In general terms, a statistic that indicates the inconsistency in test scores 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 rater. 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. See **Measurement Error**, **Observed Score**,

Standard Deviation, **True Score**.

Standardized Conditions — The aspects of the test and testing environment that are required to be the same for most test takers, to allow fair comparison of their scores. Exceptions to standardized conditions are made for test takers in need of accommodations or modifications. See **Accommodation**, **Modification**.

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 states the official position of these organizations about the appropriate ways to develop, use, and evaluate tests.

Studied Group — A population group for which special analyses of test performance are conducted because of concerns about fairness of the test to members of that group. See **Population Group**.

Subscore — A score derived from a subset of the items in a test.

Summative Evaluation — Testing a student at the completion of a unit of instruction. Compare **Formative Evaluation**.

Test — A systematic sample of behavior taken to allow inferences about an individual's knowledge, skill, ability, or other attribute.

Test Center — A site where tests are administered.

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

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

Test Results — Indicators of test performance, including scores, score profiles, diagnostic information, status labels (e.g., *Proficient*), aggregated data, imputed scores and summaries of individual or group performance. See **Score**.

Test-Retest Reliability — An evaluation 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, the selection of items is not included as a source of inconsistency. 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.

Trait Scoring — A synonym for Analytic Scoring. See **Analytic Scoring**. Compare **Holistic Scoring**.

True Score — In theory, the average of the scores the test taker would have earned on all possible forms of the test (scored by all possible qualified raters, etc.). Generally not a score that is actually possible on a single administration of the test. See **Alternate Form**, **Reliability**, **Standard Error of Measurement**. Compare **Observed Score**.

User — Individual or institution making decisions on the basis of test scores.

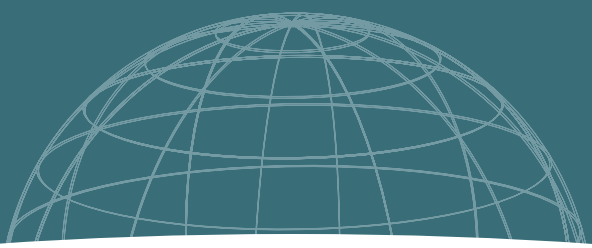
Validity — The extent to which the interpretations of scores, the claims made about test takers, and inferences and actions made on the basis of a set of scores are appropriate and justified by evidence. Validity refers to how the scores are used rather than to the test itself. Validity is a unified concept, but several aspects of validity evidence are often distinguished. Compare **Construct Validity**, **Content Validity**, **Criterion-Related Validity**.

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

Variable — An attribute that can take on different values, such as a person's test score, grade point average, height, age, etc.

Variance — A statistic that indicates the amount of variation in a set of scores, expressed in units that are the square of the units the scores are expressed in. The square root of the variance is the Standard Deviation. See **Standard Deviation**.

Weighting — (1) Specifying the relative contribution of each score, when two or more scores are combined to form a composite score. See **Composite Score**. (2) Computing statistics in a way that gives some test takers greater influence than others, usually for the purpose of estimating the statistics in some other group of test takers (i.e., other than the group for which data are available). A common use of weighting is to make a sample of test takers more accurately represent a population. See **Population**, **Sample**.



Listening. Learning. Leading.®

www.ets.org