

Technical Manual for the *Praxis*® Tests and Related Assessments

ETS®
**Professional
Educator**
PROGRAMS





Table of Contents

PREFACE.....	5
Purpose of This Manual.....	5
Audience	5
PURPOSE OF THE <i>PRAXIS</i> ® ASSESSMENTS	6
Overview.....	6
The <i>Praxis</i> ® Core Academic Skills for Educators Tests —Academic Competency	7
The <i>Praxis</i> ® Subject Assessments — Subject Knowledge and Pedagogical Knowledge Related to Teaching.....	7
The School Leadership Series Assessments	7
How the <i>Praxis</i> Assessments Address States’ Needs.....	8
ASSESSMENT DEVELOPMENT.....	9
Fairness in Test Development.....	9
Test Development Standards	9
Validity	10
Test Development Process.....	13
Review Processes.....	18
TEST ADOPTION PROCESS	19
Process Overview	19
Analysis of States’ Needs	22
Standard-Setting Studies.....	22
PSYCHOMETRIC PROPERTIES.....	25
Introduction.....	25
Test-Scoring Process.....	25
Item Analyses	26
Differential Item Functioning (DIF) Analyses	29
Test-Form Equating	31
Test Statistics	35
SCORING METHODOLOGY	38
Scoring.....	38
Scoring Methodology for Constructed-Response Items.....	38
Content Category Information.....	41
Quality Assurance Measures	41
Appropriate Score Use.....	41



SCORE REPORTING	43
Scanning/Scoring	43
Score Reports	43
Score Information for States and Institutions	43
Title II Reporting	44
APPENDIX A — THE <i>PRAXIS</i> ® JOB ANALYSES.....	46
APPENDIX B – STATISTICAL CHARACTERISTICS OF THE <i>PRAXIS</i> ® CORE ACADEMIC SKILLS FOR EDUCATORS TESTS, THE <i>PRAXIS</i> ® SUBJECT ASSESSMENTS, AND SCHOOL LEADERSHIP SERIES TESTS	47
BIBLIOGRAPHY	55

Preface

Purpose of This Manual

The purpose of the Technical Manual for the *Praxis*® Tests and Related Assessments is to explain:

- The purpose of the *Praxis*® tests
- How states use the *Praxis* tests
- The approach ETS takes in developing the *Praxis* tests
- The validity evidence supporting the *Praxis* test score use
- How states adopt the *Praxis* tests for use in their programs
- The statistical processes supporting the psychometric quality of the *Praxis* tests
- The score reporting process
- Statistical summaries of test taker performance on all of the *Praxis* tests

Audience

This manual was written for policy makers and state educators who are:

- Interested in knowing more about the *Praxis* program
- Interested in how *Praxis* relates to state licensure programs
- Interested in understanding how the *Praxis* tests are developed and scored
- Interested in the statistical characteristics of the *Praxis* tests

***Purpose of the Praxis*® Assessments**

Overview

ETS's mission is to advance quality and equity in education by providing fair and valid tests, research, and related services. In support of this mission, ETS has developed the *Praxis*® assessments. The *Praxis* tests provide states with testing tools and ancillary services that support their teacher licensure and certification process. These tools include tests of academic competency and subject-specific assessments related to teaching.

All states have an abiding interest in ensuring that teachers have the knowledge and skills needed for safe and effective practice before they receive a license. To aid states in this effort, the *Praxis* tests are designed to assess test takers' job-relevant knowledge and skills. States adopt the *Praxis* tests as one indicator that teachers have achieved a specified level of mastery of academic skills, subject area knowledge, and pedagogical knowledge before being granted a teaching license.

Each of the *Praxis* tests reflects what practitioners in that field across the United States believe to be important for new teachers. The knowledge and skills measured by the tests are informed by this national perspective, as well as by the content standards recognized by that field. The nexus of these sources of knowledge and skills means that the *Praxis* assessments offer states the opportunity to understand if their test takers are meeting the expectations of the profession. *Praxis* test scores are portable across states and directly comparable, reinforcing interstate eligibility and mobility. A score earned by a person who takes a *Praxis* test in one state represents the same level of knowledge or skill as the same score obtained by a person who takes the same *Praxis* test in another state.

The use of *Praxis* tests by large numbers of states also means that multiple forms of each assessment are rotated throughout the testing year. This minimizes the possibility of a test taker earning a score on the test due to having had prior experience with that test form on a previous administration. This feature of test quality assurance is difficult to maintain when testing volumes are too low, which is often the case with smaller, single-state testing programs.

States also customize their selection of the *Praxis* assessments. There is frequently more than one test in a content series: mathematics, social studies, English Language Arts, etc. States are encouraged to select those *Praxis* assessments that best suit their particular needs. States also customize their passing-score requirements on the *Praxis* assessments. Each state may hold different expectations for what is needed to enter the teaching profession in that field in that state. Each state ultimately sets its own passing score, which may be different from that of another state. This interplay between interstate comparability and in-state customization distinguishes the *Praxis* licensure tests.

The *Praxis*® Core Academic Skills for Educators Tests —Academic Competency

The *Praxis* Core Academic Skills for Educators tests are designed to measure academic competency in reading, writing, and mathematics. The tests are taken on computer. Many colleges, universities, and other institutions use the results of the *Praxis* Core tests as a way of evaluating test takers for entrance into teacher education programs. Many states use the tests in conjunction with *Praxis* Subject Assessments as part of the teacher licensing process.

The *Praxis*® Subject Assessments — Subject Knowledge and Pedagogical Knowledge Related to Teaching

Some *Praxis* Subject Assessments cover general or specific content knowledge in a wide range of subjects across elementary school, middle school, or high school. Others, such as the Principles of Learning and Teaching tests, address pedagogy at varying grade levels by using a case-study approach with selected-response (SR) and constructed-response (CR) items.

States that have chosen to use one or more of the *Praxis* tests require their applicants to take the tests as part of the teacher licensure process. The test provides states with a standardized mechanism to assess whether prospective teachers have demonstrated knowledge believed to be important for safe and effective entry-level practice. In addition, some professional associations and organizations require specific *Praxis* tests as one component of their professional certification requirements.

The content matter of the *Praxis* Subject Assessments is defined and validated by educators in each subject area tested. ETS oversees intensive committee work and national job analysis surveys so that the specifications for each test are aligned with the knowledge expected of the entry-level teacher in the relevant content area. In developing test specifications, standards of professional organizations also are considered, such as the standards of the National Council of Teachers of Mathematics or the National Science Teachers Association. (A fuller description of these development processes is provided in later chapters.)

When a state is considering the adoption of one or more *Praxis* Subject Assessments, state representatives are closely involved in many phases of development. For example, teachers of that content area and faculty who prepare teachers are involved in evaluating the relevancy of the test content for entry-level teachers in that state. If a test is adopted, teachers and faculty who prepare teachers are involved in the process of recommending a passing score for the test to appropriate teacher licensure agencies. Teachers also are involved in development activities, such as writing and reviewing test items and serving on test committees. Input from educators, therefore, is vital in the development of the *Praxis* tests.

The School Leadership Series Assessments

The School Leadership Series (SLS) assessments were developed for states to use as part of the licensure process for principals, superintendents and other school leaders.

These tests reflect the most current standards on professional judgment and the experiences of educators across the country. These assessments are based on the Professional Standards for Educational Leaders (PSEL) and the input of practicing school- and district-level administrators and faculty who prepare educational leaders. As with the *Praxis* Subject Assessments, educational

leaders and faculty who prepare educational leaders recommend passing scores to licensure agencies responsible for licensing principals and superintendents.

How the *Praxis* Assessments Address States' Needs

States have always wanted to ensure that beginning teachers have the requisite knowledge and skills necessary. The *Praxis* tests provide states with the appropriate tools to make decisions about applicants for a teaching license. In this way, the *Praxis* tests meet the basic needs of state licensing agencies. But the *Praxis* tests provide more than this essential information.

Over and above the actual tests, the *Praxis* program provides states with ancillary materials that help them make decisions related to licensure. For example, when states evaluate teacher licensure assessment programs, it is important to understand the nature of each program, its objectives, and the benefits it provides to the state, teacher candidates, and any other state constituents who will use the test results to inform licensure decisions. Information to help decision makers understand the critical issues associated with teacher assessment programs is available on the *Praxis* website at http://www.ets.org/praxis/states_agencies. In addition, ETS has developed a guide, *Proper Use of The Praxis Series® and Related Assessments*, to help decision makers address those critical issues. Some of the topics in the guide are:

- How the *Praxis* tests align with state and national content standards.
- How the *Praxis* tests complement existing state infrastructures for teacher licensure.
- How the *Praxis* tests are appropriate for both traditional and alternate-route candidates.

States also want to ensure that their applicants' needs are being met. To that end, the *Praxis* program has available many helpful test preparation tools. These materials take many forms:

- Free Study Companions, available online for download, including test specifications, sample questions with answers and explanations, and study tips and strategies.
- Interactive Practice Tests that simulate the computer-delivered test experience and allow test takers to practice answering authentic test questions and review answers with explanations
- A computer-delivered testing demonstration and videos on Strategies for Success and What to Expect on the Day of Your Computer-delivered Test
- Live and pre-recorded webinars detailing how to develop an effective study plan

Finally, states have a strong interest in supporting their Educator Preparation Programs. The *Praxis* program has made available the ETS Data Manager for the *Praxis* tests, a collection of services related to *Praxis* score reporting and analysis. These services are designed to allow state agencies, national organizations, and institutions to receive and/or analyze *Praxis* test results. Offered services include Quick and Custom Analytical Reports, Test-taker Score Reports and Test-taker Score Reports via Web Service. Each year, institutions also receive annual summary reports of their *Praxis* test takers' scores. The *Praxis* tests also offer an additional Title II Reporting Service to institutions of higher education to help them satisfy federal reporting requirements.

Assessment Development

Fairness in Test Development

ETS is committed to assuring that its tests are of the highest quality and as free from bias as possible. All ETS products and services—including individual test items, tests, instructional materials, and publications—are evaluated during development so that they are not offensive or controversial; do not reinforce stereotypical views of any group; are free of racial, ethnic, gender, socioeconomic, or other forms of bias; and are free of content believed to be inappropriate or derogatory toward any group.

For more explicit guidelines used in item development and review, please see the [ETS Fairness Guidelines](#).

Test Development Standards

During the *Praxis*® test development process, the program follows the strict guidelines detailed in [Standards for Educational and Psychological Testing](#) (AERA, APA, NCME, 2014):

- Define clearly the purpose of the test and the claims one wants to make about the test takers
- Develop and conduct job analysis/content validation surveys to confirm domains of knowledge to be tested
- Develop test specifications and test blueprints consistent with the purpose of the test and the domains of knowledge supported by the job analysis
- Develop specifications for item types and numbers of items needed to adequately sample the domains of knowledge supported by the job analysis survey
- Develop test items that provide evidence of the measurable-behavior indicators detailed in the test specifications
- Review test items and assembled test forms so that each item has a single best defensible answer and assesses content that is job relevant
- Review test items and assembled forms for potential fairness or bias concerns, overlap, and cueing, revising or replacing items as needed to meet standards¹.

¹ Cueing refers to an item that points to or contains the answer to another question. For example, an item may ask, “Which numbers in this list are prime numbers?” A second item may say, “The first prime numbers are... What is the next prime number in the sequence?” In this case, the second question may contain the answer to the first question.

Validity

The Nature of Validity Evidence

A test is developed to fulfill one or more intended uses. The reason for developing a test is fueled, in part, by the expectation that the test will provide information about the test taker's knowledge and/or skill that:

- May not be readily available from other sources
- May be too difficult or expensive to obtain from other sources
- May not be determined as accurately or equitably from other sources.

But regardless of why a test is developed, evidence must show that the test measures what it was intended to measure and that the meaning and interpretation of the test scores are consistent with each intended use. Herein lies the basic concept of validity: the degree to which evidence (rational, logical, and/or empirical) supports the intended interpretation of test scores for the proposed purpose ([Standards for Educational and Psychological Testing](#)).

A test developed to inform licensure² decisions is intended to convey the extent to which the test taker (candidate for the credential) has a sufficient level of knowledge and/or skills to perform important occupational activities in a safe and effective manner ([Standards for Educational and Psychological Testing](#)). “Licensure is designed to protect citizens from mental, physical, or economic harm that could be caused by practitioners who may not be sufficiently competent to enter the profession” (Schmitt, 1995). A licensure test is often included in the larger licensure process—which typically includes educational and experiential requirements—because it represents a standardized, uniform opportunity to determine if a test taker has acquired and can demonstrate adequate command of a domain of knowledge and/or skills that the profession has defined as being important or necessary to be considered qualified to enter the profession.

The main source of validity evidence for licensure tests comes from the alignment between what the profession defines as knowledge and/or skills important for safe and effective practice and the content included on the test ([Standards for Educational and Psychological Testing](#)). The knowledge and/or skills that the test requires the test taker to demonstrate must be justified as being important for safe and effective practice and needed at the time of entry into the profession. “The content domain to be covered by a credentialing test should be defined and clearly justified in terms of the importance of the content for credential-worthy performance in an occupation or profession” ([Standards for Educational and Psychological Testing](#), p. 181). A licensure test, however, should not be expected to cover all occupationally relevant knowledge and/or skills; it is only the subset of this that is most directly connected to safe and effective practice at the time of entry into the profession ([Standards for Educational and Psychological Testing](#)).

The link forged between occupational content and test content is based on expert judgment by practitioners and other stakeholders in the profession who may have an informed perspective about requisite occupational knowledge and/or skills. Processes for gathering and analyzing content-related validity evidence to support the relevance and importance of knowledge and/or skills

² Licensure and certification tests are referred to as credentialing tests by the *Standards for Educational and Psychological Testing* (2014). Unless quoted from the Standards, we use the term “licensure.”

measured by the licensure test are important for designing the test and monitoring the continued applicability of the test in the licensure process.

Within the test development cycle, the items in the *Praxis* Core Academic Skills for Educators tests and *Praxis* Subject Assessments are developed using an evidence-centered design process (ECD) that further supports the intended uses of the tests.³ Evidence-centered design is a construct-centered approach to developing tests that begins by identifying the knowledge and skills to be assessed (see “Content-related Validity Evidence” on page 11). Building on this information, test developers then work with a National Advisory Committee (NAC), asking what factors would reveal those constructs and, finally, what tasks elicit those behaviors. This design framework, by its very nature, makes clear the relationships among the inferences that the assessor wants to make, the knowledge and behaviors that need to be observed to provide evidence for those inferences, and the features of situations or tasks that evoke that evidence. Thus, the nature of the construct guides not only the selection or construction of relevant items but also the development of scoring criteria and rubrics. In sum, test items follow these three ECD stages: a) defining the claims to be made, b) defining the evidence to be collected, and c) designing the tasks to be administered.

Content-related Validity Evidence.

The [*Standards for Educational and Psychological Testing*](#) makes it clear that a systematic examination, or job analysis, needs to be performed to provide content-related validity evidence for the validity of a licensure test: “Typically, some form of job or practice analysis provides the primary basis for defining the content domain [of the credentialing test]” (p. 182). A job analysis refers to a variety of systematic procedures designed to provide a description of occupational tasks/responsibilities and/or the knowledge, skills, and abilities believed necessary to perform those tasks/responsibilities. The design of *Praxis* tests uses the following process:

- A review of available professional literature and disciplinary (content) standards to develop a draft domain of knowledge and/or skills
- Meetings with NAC experts to review and revise the draft domain
- A survey of the profession to confirm the importance of the committee-revised domain (see, for example, Knapp and Knapp, 1995; Raymond, 2001; Tannenbaum and Rosenfeld, 1994).
- Development of test specifications or blueprints based on the guidance provided by the NAC and the larger survey of educators.

For each new and revised *Praxis* test, a NAC is assembled that

- is diverse with respect to race, ethnicity, and gender
- is representative of different practice settings, grade levels, and geographic regions
- reflects different professional perspectives.

Such diversity and representation reinforces the development of domain knowledge and/or skills that is applicable across the profession. The involvement of various subgroups of experts also is part of the process of developing a test that is considered fair and reasonable to subgroups of

³ Williamson, D.M., Almond, R.G., and Mislevy, R.J. (2004). Evidence-centered design for certification and licensure. *CLEAR Exam Review*, Volume XV, Number 2, 14–18.

practitioners and test takers. Broadly, the committee's charge is to review and revise the draft domain so that it adequately defines the knowledge and/or skills important for safe and effective entry-level practice.

An online survey of the draft content domain job analysis is conducted to obtain independent judgments of the importance of the knowledge and/or skills defined by the NAC. The survey is an opportunity to collect input from a large, diverse (defined, for example, by race and ethnicity, gender, geographic region, practice setting) group of practitioners and other relevant stakeholders with an informed occupational perspective. The purpose of the survey is to determine which specific knowledge and/or skills are verified to be important and needed when entering the profession.

Basic analyses (means and standard deviations) are conducted to summarize and interpret responses. Analyses are conducted for the total group of respondents and for subgroups of respondents. The purpose of these analyses is to identify those knowledge and/or skill statements that have been judged to be most important for entering teachers; the subgroup analysis is used to identify content that may not be similarly valued by different groups of educators.

The input from the NAC and the survey are used to inform the development of test content specifications that serve as the blueprint for formal test development. It is this alignment between content-related validity evidence and test content specifications and, ultimately, between test content specifications and test items, that serves to reinforce the appropriate use of test scores for licensure purposes.

A list of ETS job analyses can be found in “Appendix A – The Praxis Series Job Analyses.”

Validity Maintenance

The content covered by a *Praxis* Core Academic Skills for Educators test, *Praxis* Subject Assessment, or School Leadership Series licensure test is evaluated on a regular basis so that it accurately reflects the current state of knowledge and/or skill requirements of that profession. Each year, the test specifications are reviewed by members of a subject-specific Standing Committee to determine if the content being measured continues to reflect relevant and important knowledge and skills recognized by the field. Particular attention is given to test specifications in disciplines when national standards are revised (e.g., release of the 2017 International Literacy Association standards). If significant changes in the field are noted, the design process discussed above is repeated to revise the test.

Test Development Process

Praxis tests and related materials follow a rigorous development process, as outlined below and in Figure 1:

- Research professional standards to develop the claims to be supported by the test scores and draft the content domain aligned with the standards, and gather content-related validity evidence to support the content domain (i.e., job analysis).
- Convene a National Advisory Committee (NAC) of subject matter experts to develop test specification
- Recruit subject-matter experts, which include practitioners in the field as well as professors, who teach the potential test takers and understand the job defined in the job analysis, to write items for the test.
- Develop sufficient numbers of test items to form a pool from which parallel forms can be assembled.
- Review the items developed by trained writers, applying and documenting [*ETS Standards for Quality and Fairness*](#) (2014) and editorial guidelines. Each item is independently reviewed by multiple reviewers who have the content expertise to judge the accuracy of the items. Note that external reviews are required at the form level, not at the item level.
- Prepare the approved test items for use in publications or tests.
- Send assembled test(s) to appropriate content experts for a final validation of the match to specifications, importance to the job, and accuracy of the correct response.
- Perform final quality-control checks according to the program's standard operating procedures to ensure assembled test(s) are ready to be administered.
- Administer a pilot test if it is included in the development plan.
- Analyze and review test data from the pilot or first administration to verify that items are functioning as intended and present no concerns about the intended answers or impact on subgroups.

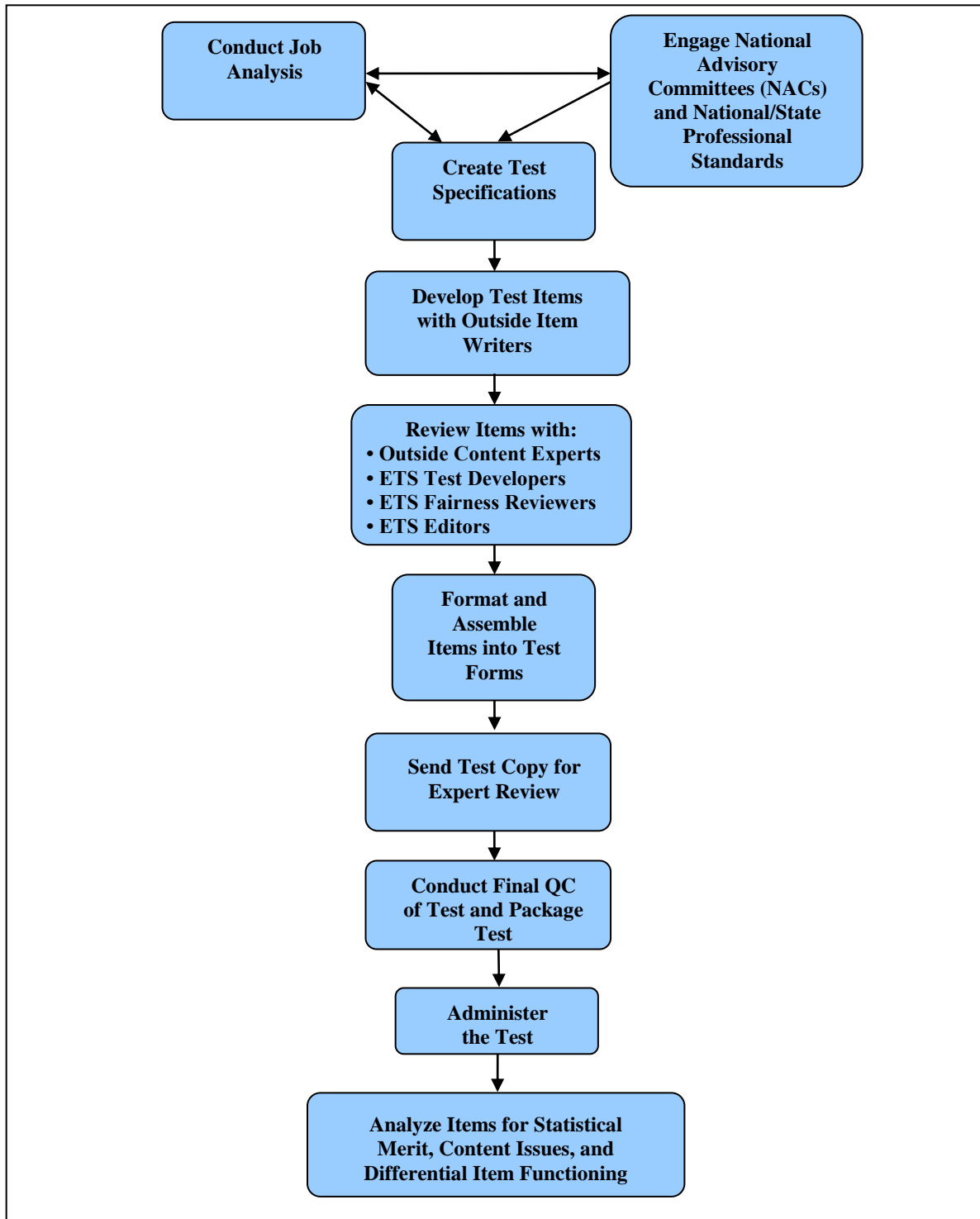


Figure 1: Test Development Process

This section details each of the steps shown in Figure 1.

Conduct Job Analysis.

Job analysis is a systematic process, the goal of which is to determine the knowledge and/or skills important for safe and effective entry-level practice. As explained in “Validity,” a job analysis is the primary source of validity evidence for licensure tests ([Standards for Educational and Psychological Testing](#)). One phase of the job analysis is the construction of a domain of job-related knowledge and/or skills. This is done with the assistance of a NAC. The NAC-approved domain is administered as a survey to a large sample of teachers and college faculty for verification of the judged importance of the knowledge and/or skills for entry-level practice. The outcomes of the survey are then used by the NAC to develop test content specifications. A list of job analyses related to the *Praxis* tests is on page 46.

Engage National Advisory Committees (NACs) and National/State Professional Standards

The NAC is a group of approximately 15 experts in the field—practicing teachers, teacher educators, and administrators—nominated by state departments of education, professional organizations, deans, superintendents, and colleagues to participate in developing job-related test content specifications. The specifications are necessary to support the validity of licensure test score use ([Standards for Educational and Psychological Testing](#)).

The NAC has two primary roles during the test development process. First, the NAC is tasked with reviewing a draft domain of knowledge and/or skill statements believed to be important for entry-level practice. The draft reflects the current state of the profession as defined by state and/or national standards. The NAC is asked to revise the draft so that it adequately reflects what the NAC considers to be important for entry-level practice. The NAC members are asked to consider each draft knowledge and/or skill statement in relation to three criteria:

1. The importance of the knowledge and/or skill for safe and effective practice
2. Whether the knowledge and/or skill is needed upon entry into the profession
3. Whether the knowledge and/or skill statement is clear and understandable.

For a knowledge and/or skill statement to be included in the domain description it must satisfy each criterion; that is, a knowledge and/or skill statement must be *important* and *needed upon entry*, and be *clear and understandable*. Not all statements meet the three criteria; those that do not are deleted from the domain description.

Second, the NAC works with ETS test developers to translate the content domain into test specifications described below.

Create Test Specifications

Test specifications are documents that inform stakeholders of the essential features of tests. These features include:

- A statement of the purpose of the test and a description of the test takers
- The major categories of knowledge and/or skills covered by the test and a description of the specific knowledge and/or skills that define each category; the proportion that each major category contributes to the overall test; and the length of the test
- The kinds of items on the test
- How the test will comply with [*ETS Standards for Quality and Fairness*](#).

The test specifications also are used to direct the work of item writers by providing explicit guidelines about the types of items needed and the specific knowledge and/or skills that each item needs to measure.

Develop Test Items with Outside Item Writers

Content experts, external to ETS, are recruited to develop test items. The experts are educators who know the domains of knowledge to be tested and are adept at using the complexities and nuances of language to write items at various difficulty levels. They write items that match the behavioral objectives stated in the test specifications and their items are written to provide sufficient evidence that the test taker is competent to begin practice.

Review Items with Outside Content Experts, ETS Test Developers, ETS Fairness Reviewers, ETS Editors

The outside review of items is an essential step in the validity chain of evidence required by good test development practice. All items for use on a *Praxis* test are vetted by practicing teachers for importance and job relevance and by other content experts for match to specifications and correctness of intended response.

In addition to the reviews of outside experts, all items used on a test are reviewed by ETS test developers, fairness reviewers, and editors. Changes to items are documented and discussed often before the final version is used on a test.

Format and Assemble Items into Test Forms

ETS test developers assemble a test form using items that have been reviewed and approved by edit, fairness reviewers, and content experts. A preview of the items selected to be used in a form is then generated for test developers to check for quality. Any changes to items are documented in the electronic assembly unit record. The test coordinator then confirms all changes have been made correctly and verifies that the standards documented in the program's Standard Operating Procedures (SOPs) have been met.

Send Test Copy for Expert Review

Before a test is certified by test developers and the test coordinator as ready to be administered, it receives a content review to verify that every item has a single best answer, which can be defended, and that no item has more than one possible key. The reviewer must understand the purpose of the test and be prepared to challenge the use of any item that is not important to the job of the beginning practitioner or is not a match to the test specifications.

Conduct Final QC of Test and Package Test

When content reviews of a particular test form have been completed, test developers perform multiple checks of the reviewers' keys against the official key and address each reviewer's comment. Once test developers deem the test ready, test coordinators then check that all steps specified in the SOPs have been followed. They must certify that the test is ready for packaging; that is, the test is ready to be administered to test takers.

Administer the Test

When the decision to develop a new form for a particular test title is made, it also is decided which of the *Praxis* general administration dates will be most advantageous for introducing the new form. This decision is entered in the Test Form Schedule, which contains specific information about test dates, make-up dates, and forms administered on each testing date for each of the *Praxis* test titles.

Analyze Items for Statistical Merit, Content Issues, and Differential Item Functioning

In the week following an administration, test developers receive the psychometrician's preliminary item analysis (PIA). In addition to item analysis graphs (see *Item Analyses*), PIA output contains a list of flagged items that test developers must examine to verify that each has a single best answer. Test developers consult with a content expert on these flagged items and document the decisions to score (or not to score) the items in a standard report prepared by the statisticians. Test developers must provide a rationale for the best answer to each flagged item as well as an explanation as to why certain flagged distracters are not keys.

If it is decided not score an item, a Problem Item Notice (PIN) is issued and distributed. The distribution of a PIN triggers actions in the Statistical Analysis, Assessment Development, and Score Key Management organizations. As a result, items in databases may need to be revised and number of items used to compute and report scores, adjusted.

If there is sufficient test taker volume, Differential Item Functioning (DIF) analyses are run on a new test form to determine if subgroup differences in performance may be due to factors other than the abilities the test is intended to measure. These procedures are described more fully in "Differential Item Functioning (DIF) Analyses" on page 29, and in Holland and Wainer (1993). A DIF panel of content experts decides if items with statistically high levels of DIF (C-DIF) should be dropped from scoring. If that is the case, test developers must prepare a do-not-score PIN and close a report using test creation software. Test developers are responsible for ensuring that C-DIF items are not used in future editions of the test.



Review Processes

ETS has strict, formal review processes and guidelines. All ETS licensure tests and other products undergo multistage, rigorous, formal reviews to verify that they adhere to ETS's fairness guidelines that are set forth in three publications:

ETS Standards for Quality and Fairness

Every test that ETS produces must meet the exacting criteria of the [*ETS Standards for Quality and Fairness*](#). These standards reflect a commitment to producing fair, valid, and reliable tests. The criteria are applied to all ETS-administered programs, and compliance with them has the highest priority among the ETS officers, Board of Trustees, and staff. Additionally, the ETS Office of Professional Standards Compliance audits each ETS testing program to ensure its adherence to the [*ETS Standards for Quality and Fairness*](#).

In addition to complying with the ETS quality standards, ETS develops and administers tests that comply with the [*Standards for Educational and Psychological Testing*](#) (2014) and [*The Code of Fair Testing Practices in Education*](#).

ETS Fairness Review

The [*ETS Fairness Guidelines*](#) identifies aspects of test items that might hinder people in various groups from performing at optimal levels. Fairness reviews are conducted by specially trained reviewers.

Test Adoption Process

Process Overview

The Praxis® Core Academic Skills for Educators Tests

Teacher Licensure. The *Praxis* Core Academic Skills for Educators tests may be used by the licensing body or agency within a state for teacher licensure decisions. The *Praxis* program suggests that before adopting a test, the licensing body or agency reviews the test specifications to confirm that the content covered on the test is consistent with state standards and with expectations of what the state’s teachers should know and be able to do. The licensing body or agency also must establish a passing standard or “cut score.” ETS conducted a multistate standard-setting study for the *Praxis* Core and provided the results to the licensing body or agency to inform its decision.

Entrance into Teacher Preparation Programs. These tests also may be used by institutions of higher education to identify students with sufficient reading, writing, and mathematics skills to enter a teacher preparation program. If an institution is in a state that has authorized the use of the *Praxis* Core tests for teacher licensure and has set a passing score, the institution may use the same minimum score requirement for entrance into its program. Even so, institutions are encouraged to use other student qualifications, in addition to the *Praxis* Core scores, when making final entrance decisions.

If an institution of higher education is in a state that has not authorized use of the *Praxis* Core tests for teacher licensure, the institution should review the test specifications to confirm that the skills covered are important prerequisites for entrance into the program; it also will need to establish a minimum score for entrance. These institutions are encouraged to use additional student qualifications when making final entrance decisions.

The Praxis® Subject Assessments

Teacher Licensure. The *Praxis* Subject Assessments may be used by the licensing body or agency within a state for teacher licensure decisions. This includes test takers who seek to enter the profession via a traditional or state-recognized alternate route as well as those currently teaching on a provisional or emergency certificate who are seeking regular licensure status. The licensing body or agency also must establish passing standards or “cut scores.” ETS conducts multistate standard-setting studies for the *Praxis* Subject tests and provides the results to the licensing body or agency to inform its decision.

Program Quality Evaluation. Institutions of higher education may want to use *Praxis* Subject Assessments scores as one criterion to judge the quality of their teacher preparation programs. The *Praxis* program recommends that such institutions first review the test’s specifications to confirm alignment between the test content and the content covered by the preparation program.

Entrance into Student Teaching. Institutions of higher education may want to use *Praxis* Subject Assessments content scores as one criterion for permitting students to move on to the clinical portion of their program: the student teaching phase. This use of the *Praxis* Subject Assessment is often

based on the argument that a student teacher should have a level of content knowledge comparable to that of a teacher who has just entered the profession. This argument does not apply to pedagogical skills or knowledge, so the *Praxis*® tests that only focus on pedagogical knowledge (i.e., the Principles of Learning and Teaching set of assessments) should not be used as prerequisites for student teaching.

There are three scenarios involving the use of *Praxis* content assessments for entrance into student teaching: (1) The state requires that all content-based requirements for licensure be completed before student teaching is permitted; (2) The state requires the identified *Praxis* Subject Assessments content test for licensure, but not as a prerequisite for student teaching; and (3) The state requires the identified *Praxis* content test neither for licensure nor as a prerequisite for student teaching.

If an institution is in a state that uses the identified *Praxis* content assessment for licensure, the state may also require test takers to meet its content-based licensure requirements before being permitted to student teach. In this case, additional validity evidence on the part of the program may not be necessary, as the state, through its adoption of the test for licensure purposes, has accepted that the test's content is appropriate; set a schedule for when content-based licensure requirements are to be met; and already established the passing scores needed to meet its requirements.

The following summarizes this process:

IF...	THEN...
a state requires content-based licensure before student teaching is allowed	Additional validity evidence is not necessary if the state: <ul style="list-style-type: none">• Accepts the <i>Praxis</i> Subject Assessment as valid• Sets a schedule for meeting content-based licensure requirements• Establishes passing scores to meet requirements.

If an institution, but not the state, requires that students meet the content-based licensure requirement before being permitted to student teach, and the state requires the use of the identified *Praxis* content test for teacher licensure, the institution should review the test specifications to confirm that the content covered is a necessary prerequisite for entrance into student teaching and that the curriculum which students were exposed to covered that content.

The following summarizes this process:

IF...	THEN...
an institution, but not the state, requires content-based licensure before student teaching is allowed	the institution should review test specifications to confirm that the content is necessary for student teaching and that students were exposed to the curriculum that covers the appropriate content.

AND

the state requires the use of a *Praxis* Subject Assessment content test for licensure

Institutions may use the state-determined licensure passing standard as its minimum score for entrance into student teaching or they may elect to set their own minimum scores; either way, they are encouraged to use other student qualifications, in addition to the *Praxis* content scores, when making final decisions about who may teach.

If an institution of higher education wants to use the *Praxis* Subject Assessments but is in a state that has not adopted the identified subject test for teacher licensure, that institution should review the test specifications to confirm that the content covered on the test is a necessary prerequisite for entrance into student teaching and the curriculum which students were exposed to covered that content. Institutions also will need to establish a minimum score for entrance. They are encouraged to use other student qualifications, in addition to the *Praxis* content scores, when making final decisions about who may student teach.

The following summarizes this process:

IF...	THEN...
an institution wants to use the <i>Praxis</i> Subject Assessments in a state that has not authorized the content assessment for licensure	that institution should review test specifications to confirm that the content is necessary for student teaching and that students were exposed to the curriculum that covers the appropriate content.

AND

the state requires use of a *Praxis* content test for licensure

Entrance into Graduate-level Teacher Programs. Graduate-level teacher programs most often focus on providing additional or advanced pedagogical skills. These programs do not typically focus on content knowledge itself. Because of this, such programs expect students to enter with sufficient levels of content knowledge. In states that use *Praxis* Subject Assessments for licensure, sufficient content knowledge may be defined as the test taker's having met or exceeded the state's passing score for the content assessment. In this case, the program may not need to provide additional evidence of validity because the state, by adopting the test for licensure purposes, has accepted that the test content is appropriate.

However, if a graduate-level program is in a state that has not adopted the subject test, that program should review the test specifications to confirm that the content is a necessary prerequisite for entrance into the program. The program also must establish a minimum score for entrance and is encouraged to use other student qualifications, in addition to the test scores, when making final entrance decisions.

Furthermore, the test should not be used to rank test takers for admission to graduate school.

Analysis of States' Needs

ETS works directly with individual state and/or agency clients or potential clients to identify their licensure testing needs and to help the licensing authority establish a testing program that meets those needs. ETS probes for details regarding test content and format preferences and shares information on existing tests that may meet client needs. Clients often assemble small groups of stakeholders to review sample test forms and informational materials about available tests. The stakeholder group provides feedback to the client state or agency regarding the suitability of the test assessments. When a state decides that a test may meet its needs, ETS will work with the state to help it establish a passing score.

Standard-Setting Studies

To support the decision-making process for education agencies establishing a passing score (cut score) for a new or revised *Praxis* test, research staff from ETS designs and conducts multistate standard-setting studies. Each study provides a recommended passing score, which represents the combined judgments of a group of experienced educators. ETS provides a recommended passing score from the multistate standard-setting study to education agencies. In each state, the department of education, the board of education, or a designated educator licensure board is responsible for

establishing the operational passing score in accordance with applicable regulations. *ETS does not set passing scores*; that is the licensing agencies' responsibility.

Standard-setting methods are selected based on the characteristics of the particular *Praxis* test. Typically, a modified Angoff method is used for selected-response (SR) items and an extended Angoff method is used for constructed-response (CR) items. For *Praxis* tests that include both SR and CR items, both standard-setting methods are used. One or more ETS standard-setting specialists conduct and facilitate each standard-setting study.

Panel Formation

Standard-setting studies provide recommended passing scores, which represent the combined judgments of a group of experienced educators. For multistate studies, states (licensing agencies) nominate recommended panelists with (a) experience as either teachers of the particular subject area or college faculty who prepare teachers in the subject area and (b) familiarity with the knowledge and skills required of beginning teachers. ETS selects panelists to represent the diversity (race/ethnicity, gender, geographic setting, etc.) of the teacher population. Each panel includes approximately 20 educators, the majority of whom are practicing, licensed teachers in the content area covered by the test.

Typical Standard Setting Methods

For SR items, a probability-based, modified Angoff method typically is used. In this approach, for each SR item a panelist decides on the likelihood (probability or chance) that a just qualified candidate (JQC) would answer it correctly. Panelists make their judgments using the following rating scale: 0, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95, 1. The lower the value, the less likely it is that a JQC would answer the question correctly, because the question is difficult for the JQC. The higher the value, the more likely it is that a JQC would answer the question correctly. Multiple rounds of judgments are collected, with panelist discussion between rounds. A panelist's judgments are summed across SR items to calculate his individual passing score; panelists' passing scores are averaged to produce the panel's recommended passing score.

For CR items, an extended Angoff method typically is used. In this approach, for each CR item, a panelist decides on the assigned score value that would most likely be earned by a JQC. The basic process that each panelist followed is first to review the description of the JQC and then to review the item and the rubric for that item. The rubric for a CR item defines holistically the quality of the evidence that would merit a response earning a particular score. During this review, each panelist independently considers the level of knowledge/skill required to respond to the item and the features of a response that would earn particular scores, as defined by the rubric. Multiple rounds of judgments are collected, with panelist discussion between rounds. As with the method used for SR items, a panelist's judgments are summed across CR items to calculate his individual passing score; panelists' passing scores are averaged to produce the panel's recommended passing score.

For *Praxis* tests that include both SR and CR items, both methods are used and the intermediate results for the SR items and for the CR items are combined, according to the design of the test, to calculate the recommended passing score.

Standard-Setting Reports

Approximately four weeks after the standard-setting study is completed, each participating state receives a study report. For each multistate study, a technical report is produced that describes the content and format of the test, the standard-setting processes and methods, and the results of the



standard-setting study. The report also includes information about the conditional standard error of measurement for the passing score recommendation. Each state may want to consider the information from the multistate study but also other sources of information when setting the final passing score.

Psychometric Properties

Introduction

ETS' Statistical Analysis division has developed procedures designed to support the development of valid and reliable test scores for the *Praxis*® program. The item and test statistics are produced by software developed at ETS to provide rigorously tested routines for both classical and Item Response Theory (IRT) analyses.

The psychometric procedures explained in this section follow well-established, relevant standards in [*Standards for Educational and Psychological Testing*](#) (2014) and the [*ETS Standards for Quality and Fairness*](#) (2014). They are used extensively in the *Praxis* program and are accepted by the psychometric community at large.

As discussed in the Assessment Development section, every *Praxis* test has a set of test specifications that is used to create versions of each test, called test forms. Each test form has a unique combination of individual test items. The data for the psychometric procedures described below are the test taker item responses collected when the test form is administered, most often by using the item responses from the first use of a test form.

Test-Scoring Process

The *Praxis* tests are administered nationwide in several computer-based test administrations each year. They also are given regularly at computer-based test centers. The following is an overview of the test-scoring process:

- When a new SR form is introduced, a Preliminary Item Analysis (PIA) of the test items is completed within one week following the administration. Items are evaluated statistically to confirm that they perform as intended in measuring the desired knowledge and skills for beginning teachers.

For tests that include CR items, ratings by two independent scorers are combined to yield a total score for each test question.

- A Differential Item Functioning (DIF) Analysis is conducted to determine that the test questions meet ETS's standards for fairness. DIF analyses compare the performance of subgroups of test takers on each item. For example, the responses of male and female, or Hispanic and White subgroups might be compared.

Items that show very high DIF statistics are reviewed by a fairness panel of content experts, which often include representatives of the subgroups used in the analysis. The fairness panel decides if a test takers' performance on any item is influenced by factors not related to the construct being measured by the test. Such items are then excluded from the test scoring. A more detailed account of the DIF procedures followed by the *Praxis* program are provided in "Differential Item Functioning (DIF) Analyses" on page 29, and are described at length in Holland and Wainer's (1993) text.

- Test developers consult with content experts or content advisory committees to determine whether all items in new test forms meet ETS's standards for quality and fairness. Their consultations are completed within days after the administration of the test.
- Statistical equating and scaling is performed on each new test approximately two weeks after the test administration window has been completed.
- Scores are sent to test takers and institutions of higher education two to three weeks after the test administration window has closed.

A Final Item Analysis (FIA) report is completed six to eight weeks after the test administration. The final item-level statistical data is provided to test developers to assist them in the construction of future forms of the test.

Item Analyses

Classical Item Analyses

Following the administration of a new test form, but before scores are reported, a PIA for all SR items is carried out to provide information to assist content experts and test developers in their review of the items. They inspect each item, using the item statistics to detect possible ambiguities in the way the items were written, keying errors, or other flaws. Items that do not meet ETS's quality standards can be excluded from scoring before the test scores are reported.

Information from PIA is typically replaced by FIA statistics if a sufficient number of test takers have completed the test to permit accurate estimates of item characteristics. These final statistics are used for assembling new forms of the test. However, some *Praxis* tests are taken only by a small number of test takers. For these tests, FIAs are calculated using data accumulated over several test administrations.

Preliminary and final analyses include both graphical and numerical information to provide a comprehensive visual impression of how an item is performing. These data are subsequently sent to *Praxis* test developers, who retain them for future reference. An example of an item analysis graph of an SR item is presented in Figure 2.

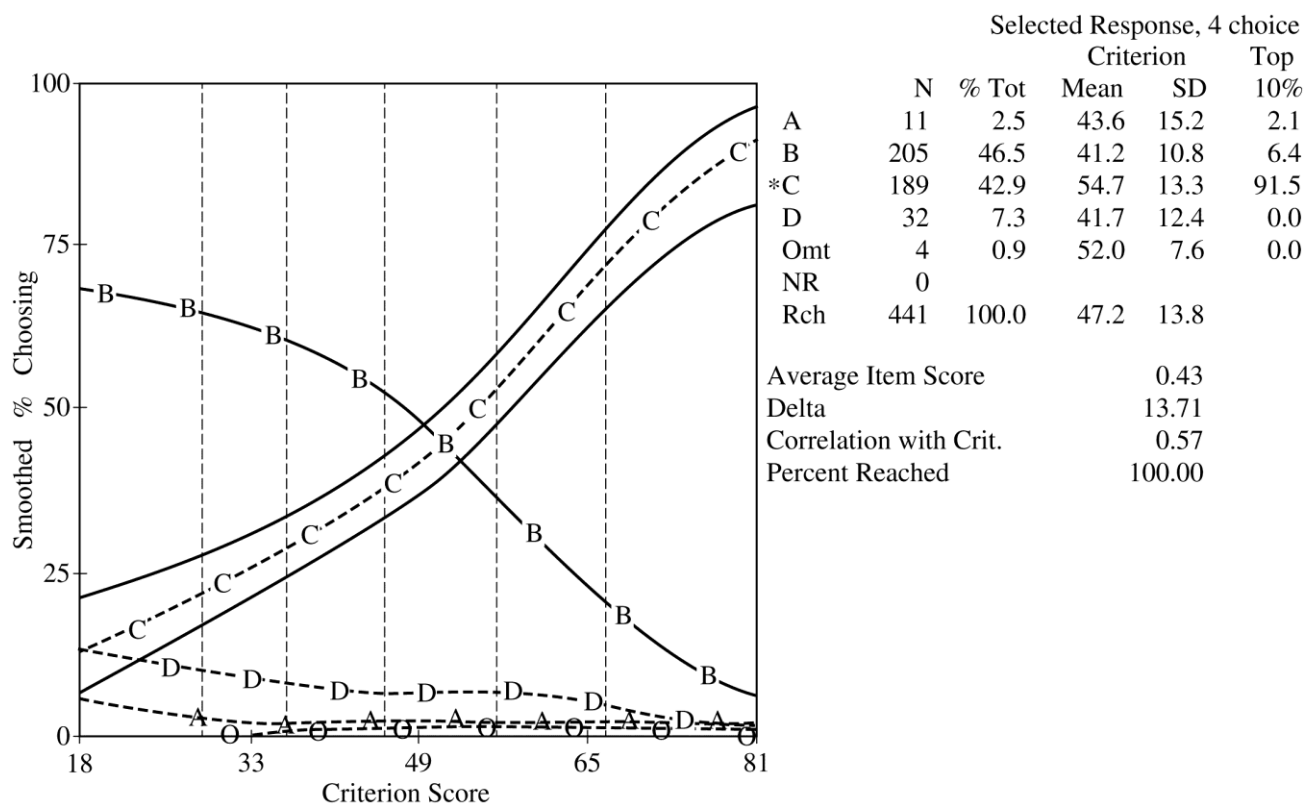


Figure 2. Example of an item analysis graph for an SR item

In this example of an SR item with four options, the percentage of test takers choosing each response choice (A–D) and omitting the item (Omt) is plotted against their performance on the criterion score of the test. In this case the criterion is the total number of correct responses. Vertical dashed lines are included to identify the 20th, 40th, 60th, and 80th percentiles of the total score distribution, and 90-percent confidence bands are plotted around the smoothed plot of the correct response (C). The small table to the right of the plot presents summary statistics for the item:

- For each response option, the table shows the count and percent of test takers who chose the option, the criterion score mean and standard deviation of respondents, and the percent of respondents with scores in the top ten percent of test takers who chose the option. The specified percentage of top scores may differ from ten percent, depending on factors such as the nature of the test and sample size.
- Four statistics are presented for the item as a whole: 1) The Average Item Score (the percent of correct responses to an item that has no penalty for guessing); 2) Delta, an index of item difficulty that has a mean of 13 and standard deviation of 4 (see footnote 6 on page 1); 3) The correlation of the item score with the criterion score. (For an SR item this is a biserial correlation, a measure of correspondence between a normally distributed continuous variable assumed to underlie the dichotomous item's outcomes, and the criterion score); 4) the percent of test takers who reached the test item.

For CR items, both item and scorer analyses are conducted. The item analyses include distributions of scores on the item; two-way tables of rater scores before adjudication of differences between scorers; the percentage of exact and adjacent agreement; the distributions of the adjudicated scores; and the correlation between the scores awarded by each of the two scorers. For each scorer, his/her scores on each item are compared to those of all other scorers for the same set of responses.

Within one week of a new form's administration, statistical analysts deliver a PIA to test developers for each new test form. Items are flagged for reasons including but not limited to:

- Low average item scores (very difficult items)
- Low correlations with the criterion
- Possible double keys
- Possible incorrect keys.

Test developers consult with content experts or content advisory committees to determine whether each SR item flagged at PIA has a single best answer and should be used in computing test taker scores. Items found to be problematic are identified by a Problem Item Notification (PIN) document. A record of the final decision on each PINned item is signed by the test developers, the statistical coordinator, and a member of the *Praxis* program direction staff. This process verifies that flawed items are identified and removed from scoring, as necessary.

When a new test form is introduced and the number of test takers is too low to permit an accurate estimation of item characteristics, the *Praxis* program uses the Testlet design described below. This test design allows items in certain portions of the test to be pretested to determine their quality before they are used operationally.

Speededness

Occasionally, a test taker may not attempt items near the end of a test because the time limit expires before she/he can reach the final items. The extent to which this occurs on a test is called "speededness." The *Praxis* program assesses speededness using four different indices:

1. The percent of test takers who complete all items
2. The percent of test takers who complete 75 percent of the items
3. The number of items reached by 80 percent of test takers⁴
4. The variance index of speededness (i.e., the ratio of not-reached variance to total score variance).⁵

All four of these indices need not be met for a test to be considered speeded. If the statistics show that many test takers did not reach several of the items, this information can be interpreted as strong evidence that the test (or a section of a test) was speeded. However, even if all or nearly all of the

⁴ When a test taker has left a string of unanswered items at the end of a test, it is presumed that he/she did not have time to attempt them. These items are considered "not reached" for statistical purposes.

⁵ An index less than 0.15 is considered an indication that the test is not speeded, while ratios above 0.25 show that a test is clearly speeded. The variance index is defined as SNR^2 / SR^2 , where SNR^2 is the variance of the number of items not reached, and SR^2 is the variance of the total raw scores.

test takers reached all or nearly all of the items, it would be wrong to conclude, without additional information, that the test (or section) was unspeeded. Some test takers might well have answered more of the items correctly if given more time. Item statistics, such as the percent correct and the item total correlation, may help to determine whether many test takers are guessing, but the statistics could indicate that the items at the end of the test are difficult. A *Praxis* Core Academic Skills for Educators test or *Praxis* Subject Assessment will be considered speeded if more than one of the speededness indices is exceeded.

Differential Item Functioning (DIF) Analyses

DIF analysis utilizes a methodology pioneered by ETS (Dorans & Kulick, 1986; Holland & Thayer, 1988; Zwick, Donoghue, & Grima, 1993). It involves a statistical analysis of test items for evidence of differential item difficulty related to subgroup membership. The assumption underlying the DIF analysis is that groups of test takers (e.g., male/female; Hispanic/White) who score similarly overall on the test or on one of its subsections—and so are believed to have comparable overall content understanding or ability—should score similarly on individual test items.

DIF analyses are conducted during the week after each *Praxis* test administration, sample sizes permitting, to inform fairness reviews. For example, DIF analysis can be used to measure the fairness of test items at a test taker subgroup level. Statistical analysts use well-documented DIF procedures, in which two groups are matched on a criterion (usually total test score, less the item in question) and then compared to see if the item is performing similarly for both groups. For tests that assess several different content areas, the more homogeneous content areas (e.g., verbal or math content) are preferred to the raw total score as the matching criterion. The DIF statistic is expressed on a scale in which negative values indicate that the item is more difficult for members of the focal group (generally African American, Asian American, Hispanic American, Native American, or female test takers) than for matched members of the reference group (generally White or male test takers). Positive values of the DIF statistic indicate that the item is more difficult for members of the reference group than for matched members of the focal group. If sample sizes are too small to permit DIF analysis before test-score equating, they are accumulated over several test administrations until there is sufficient volume to do so, usually at the end of the testing year.

DIF analyses produce statistics describing the amount of differential item functioning for each test item as well as the statistical significance of the DIF effect. ETS's decision rules use both the degree and significance of the DIF to classify items into three categories: A (least), B, and C (most). Any items classified into category C are reviewed at a special meeting that includes staff who did not participate in the creation of the tests in question. In addition to test developers, these meetings may include at least one participant not employed by ETS and a member representing one of the ethnic minorities of the focal groups in the DIF analysis. The committee members determine if performance differences on each C item can be accounted for by item characteristics unrelated to the construct that is intended to be measured by the test. If factors unrelated to the knowledge assessed by the test are found to influence performance on an item, it is deleted from the test scoring. Moreover, items with a C DIF value are not selected for subsequent test forms unless there are exceptional circumstances (e.g., the focal group performs better than the reference group, and the content is required to meet test specifications).

In addition to the analyses described previously, ETS provides test takers with a way at the test site to submit queries about items in the tests. Every item identified as problematic by a test taker is carefully reviewed, including the documented history of the item and all relevant item statistics. Test developers, in consultation with an external expert, if needed, respond to each query. When indicated, a detailed, customized response is prepared for the test taker in a timely manner.

DIF Statistics

DIF analyses are based on the Mantel Haenszel DIF index expressed on the ETS item delta scale (MH D DIF). The MH D DIF index identifies items that are differentially more difficult for one subgroup than for another, when two mutually exclusive subgroups are matched on ability (Holland & Thayer, 1985).⁶ The matching process is performed twice: 1) using all items in the test, and then 2) after items classified as C DIF have been excluded from the total score computation. For most tests, comparable (matched) test takers are defined as having the same total raw score, where the total raw score has been refined to exclude items with high DIF (C items). The following comparisons would be analyzed (if data are available from a sufficient number of test takers who indicate that English is understood as well as or better than any other language), where the subgroup listed first is the reference group and the subgroup listed second is the focal group:

- Male/Female
- White (non-Hispanic)/African American or Black (non-Hispanic)
- White (non-Hispanic)/Hispanic
- White (non-Hispanic)/Asian American
- White (non-Hispanic)/Native American, American Indian, or Alaskan Native.

The Hispanic subgroup comprises test takers who coded:

- Mexican American or Chicano
- Puerto Rican
- Other Hispanic or Latin American.

High positive DIF values indicate that the gender or ethnic focal group performed better than the reference group. High negative DIF values show that the gender or ethnic reference group performed better than the focal group when ability levels were controlled statistically.

Thus, an MH D DIF value of zero indicates that reference and focal groups, matched on total score, performed exactly the same. An MH D DIF value of +1.00 would indicate that the focal group (compared to the matched reference group) found the item to be one delta point easier. An MH D DIF of –1.00 indicates that the focal group (compared to the matched reference group) found the item to be 1 delta point more difficult.

⁶ *Delta* (Δ) is an index of item difficulty related to the proportion of test takers answering the item correctly (i.e., the ratio of the number of people who correctly answered the item to the total number who reached the item). Delta is defined as $13 - 4z$, where z is the standard normal deviation for the area under the normal curve that corresponds to the proportion correct. Values of delta range from about 6 for very easy items to about 20 for very difficult items.

Based on the results of the DIF analysis, each item is categorized into one of three classification levels (Dorans and Holland 1993), where statistical significance is determined using $p < .05$:

A = low DIF; absolute value of MH D DIF less than 1 or not significantly different from 0,

B = moderate DIF; MH D DIF significantly different from 0, absolute value at least 1, and either
(1) absolute value less than 1.5, or
(2) not significantly greater than 1,

C = high DIF; absolute value of MH D DIF at least 1.5 and significantly greater than 1.

C-level items are referred to fairness committees for further evaluation and possible revision or removal from the test. Test developers assembling a new test form are precluded from selecting C-level items unless absolutely necessary in rare cases for content coverage.

The DIF procedures described above have been designed to detect differences in performance on an item when differences in the abilities of the reference and focal groups are controlled. However, item statistics for the subgroups as a whole also are of interest. When sample sizes permit, the most commonly analyzed subgroups are defined by gender and ethnicity.

Test-Form Equating

Overview

Each *Praxis* test comprises multiple test forms, with each containing a unique set of test questions, whether selected response, constructed response, or a combination of both. [*ETS Standards for Quality and Fairness*](#) require the use of equating methodologies when “scores on alternate forms of the same test . . . are deemed interchangeable in terms of content and statistical characteristics” (page 35), as is the case for all *Praxis* tests. Equating adjusts scores on different test forms to account for the inherent inability to produce test forms with identical degrees of difficulty, even when test-assembly processes are tight. Because equating adjusts for differences in difficulty across different *Praxis* test forms, a given scale score represents the same level of achievement for all forms of the test. Well-designed equating procedures maintain the comparability of scores for a test and thus avoid penalizing test takers who happen to encounter a selection of questions that proves to be more difficult than expected (von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004).

Scaling

To avoid confusion between the adjusted and unadjusted scores, the *Praxis* program has typically reported the adjusted scores on a score scale that makes them clearly different from the unadjusted (raw) scores. This score scale is a mathematical conversion (or scaling) of the raw scores into scaled scores with predetermined lower and upper limits. *Praxis* tests use a scaled score range of 100 to 200 for score reporting. The use of a scale common to all forms of the same test title enables the users of the test to compare scores on test forms that may differ slightly in difficulty.

When the first form of a *Praxis* test consisting only of SR items is administered for the first time, the method used to establish the reported score scale is as follows:

1. The raw score to be expected by guessing randomly at each item = C
where $C = \text{Test Length} * (1 / \text{number of SR options})$

Scaled scores at or below C are fixed at the minimum possible scaled score (usually 100).

2. The score T is defined as: $\text{Test Length} * .95$

Scaled scores corresponding to raw scores of T or higher are set to the maximum scaled score for the test (usually 200).

3. For raw scores between C and T , the scaled score, S , is defined as: $S = Ax + B$
where x is the raw score, and

$$A = (\text{Scale Maximum} - \text{Scale Minimum}) / (T - C), \text{ and}$$

$$B = (\text{Scale Maximum} - \text{Scale Minimum}) - (A * C)$$

Equating

To maintain the comparability of the reported scores for each test, for each new form of a test, following the initial scaling of the first test form, each subsequent new form of a test, after its initial administration and before scores are reported, is equated to translate raw scores on the new form to adjusted scores on the test's reporting scale. The equating procedures take into account the difficulty of the form and the relative ability of the group of test takers who took that form.

The most frequently employed equating model is the Non-Equivalent groups' Anchor Test (NEAT) design, which is used in the framework of classical test theory. *Praxis* Statistical Analysis uses this design because of its relative ease of use and applicability to a variety of test settings. This approach also has the advantage of using models that work well with small samples, a possible occurrence, for example, when a new test is introduced. In fact, it may be necessary to scale the first form of a new test and then reuse it at additional administrations until accumulated volume increases sufficiently to allow the data to be used to equate a new form using the NEAT design.

The NEAT Design

Under the NEAT or anchor test design, one set of items (e.g., Test X) is administered to one group of test takers, another set of items (e.g., Test Y) is administered to a second group of test takers, and a third set of common items (e.g., Test V) is administered to both groups (Kolen & Brennan, 2004). The common items that comprise the anchor test are chosen to be representative of the items in the total tests (Test X and Test Y) in terms of both their content and statistical properties. Anchor tests can be either internal (i.e., the common items contribute to reported scores on the test form being equated) or external (i.e., the common items are not part of the test form being equated). Both linear (e.g., Tucker and Levine) and nonlinear (e.g., equipercentile) equating methods may be used under the NEAT design. The final raw-score-to-scaled-score conversion line can be chosen based on characteristics of the anchor and total test score distributions, the reliability of the tests, and the sizes of the samples used in the analysis.

The NEAT design can be used for tests comprising SR items only, CR items only, or a combination of SR and CR items:

1. Tests containing SR items only are equated using an internal anchor test. In these cases, the anchor test includes approximately 25 percent of the items in the total test.
2. Tests containing SR items and CR items are equated using only the SR items in an internal anchor test.

The Equivalent Groups Design

For tests that have a large number of test takers per administration, an equivalent group's equating design may be employed. Two different forms are administered at the same administration: an old test form with an established raw-to-scaled score conversion and a new test form. The two forms are spiraled; that is, one half of the test takers complete the old form and the other half complete the new form. Because a large number of test takers are in effect randomly assigned to take one or the other of the spiraled test forms involved, it is assumed that the average test taker's ability in each group is equivalent. Both linear and nonlinear (e.g., direct equipercentile) equating methods may be used with this design.

The Single Group Design

In certain circumstances, such as the loss of an item found to have significant DIF, a new raw-to-scaled score conversion is required to score the form without the flawed item. In these cases, a single group of test takers that has completed all the items is selected for analysis. Two sets of test statistics are calculated: one includes all items and the other omits the flawed item(s). The raw means and standard deviations of the two are set equal, establishing an estimate of the full-length test score for each possible raw score on the new (shorter) version of the test. The original raw-to-scaled score conversion is then applied to the estimates, yielding a new conversion for the shortened form.

The Testlet Design

The current equating practices explained above are not appropriate for very low volume tests (i.e., those tests that have fewer than thirty test takers per administration). For these tests, the *Praxis* program uses the testlet model. In this model, the test is constructed of a number of item clusters (called testlets). Each testlet is assembled to proportionally represent the content specifications of the full test. One of the testlets contains unscored pretest items. All testlets are carefully evaluated by content specialists when the test is assembled. A scaling of the first form of a testlet test is conducted to establish a raw-to-scaled score conversion for its first administrations. When sufficient accumulated volume is attained, a single-group equating is performed, equating a new form, created by replacing some proportion of the test form with pretest material to the original scaled test form (see Wainer & Kiely, 1987).

An example of the testlet design is shown in Figure 3, in which:

- Shaded boxes indicate testlets containing operational (scored) items.
- Unshaded boxes indicate testlets containing unscored (pretest) items.
- Solid arrows indicate a single-group equating.
- Dashed arrows indicate a change in the structure of the test form.

This exam is composed of three testlets (Operational testlets O1, O2, and O3), along with a testlet of pretest items (P1). For scoring purposes, a scaling is carried out for the first form of the test, and single-group equating is performed for the succeeding forms. In other words, when accumulated volumes are sufficient for equating, a single-group equating is performed for the two sets of scores (first set: O1 to O3; second set: O2, O3, and P1) under the assumption that O1 and P1 are sufficiently parallel with respect to content and psychometric properties. The test form composed of three item clusters (O2, O3, and P1) is converted into the scale and used at the following administration. At this stage, P1 is renamed O4, and a different set of pretest items (P2) is added to the test. The items that had comprised O1 have now been removed from the test. This revised form of the test will now replace the original form. The same replacement of operational items with pretest items will take place again after the revised form has been used at a number of test administrations and after enough test takers have completed it to permit the equating of the next form. The same linking design is then repeated: A single-group equating is carried out for the two sets of scores (first set: O2 to O4; second set: O3, O4, and P2) under the assumption that O2 and P2 are sufficiently parallel.

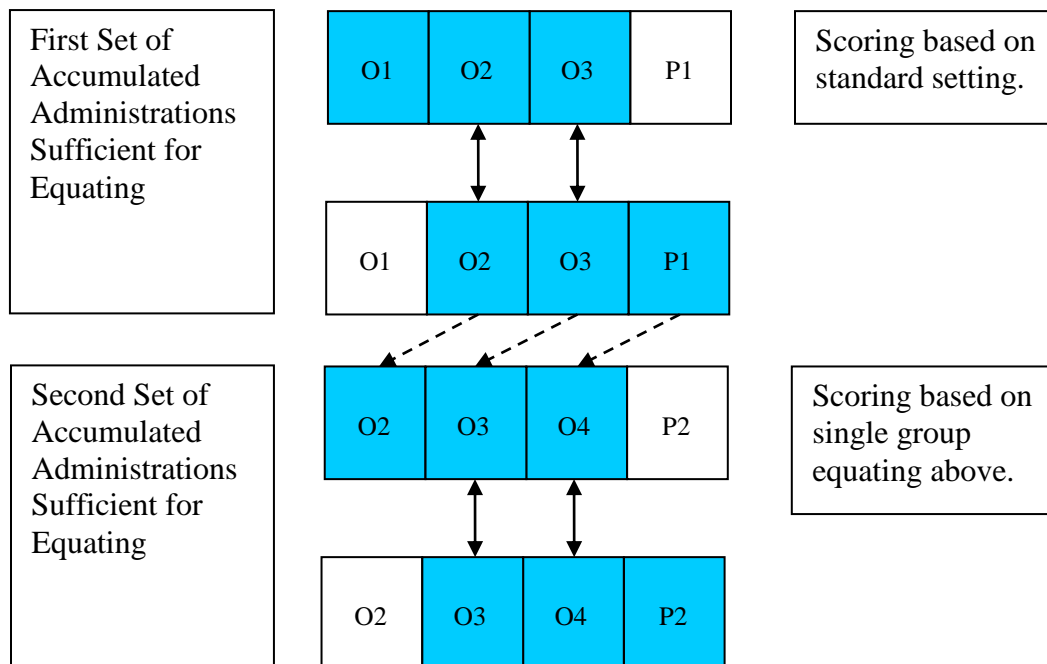


Figure 3. The Testlet Equating Design

Note: O1 to O4 = Operational items; P1 = Pretest items at Time 1; P2 = Pretest items at Time 2

Equating Methodology Summary

Because the equivalent groups equating design requires a large volume of test takers to produce dependable results, only the *Praxis* Core Academic Skills for Educators tests use this method. The smallest volume *Praxis* Subject Assessments use the testlet design. All other *Praxis* tests use the NEAT design to equate new test forms.

Test Statistics

Reliability

The reliability of a test refers to the extent to which test scores are consistent or stable. An index of reliability enables ETS to generalize beyond the specific collection of items in a particular form of a test to a larger universe consisting of all possible items that could be posed to the test taker. Because tests consist of only a sample of all possible items, any estimate of a test taker's actual capabilities will contain some amount of error. Psychometrically, reliability may be defined as the proportion of the test score variance that is due to the “true” (i.e., stable or non-random) abilities of the test takers. A person's actual (or “observed”) test score may thus be thought of as having a “true” component and an “error” component. Here, “error” is defined as the difference between the observed and true scores. Since true scores can never be known, the reliability of a set of test scores cannot be assessed directly, but only estimated.

Reliability estimates for the *Praxis* SR total, category, and equating scores are computed using the Kuder and Richardson (1937) formula 20 (KR 20). Reliability may be thought of as the proportion of test score variance that is due to true differences among the test takers with respect to the ability being measured:

$$reliability = 1 - \frac{error\ variance}{total\ variance} ,$$

If the test is not highly speeded, the KR 20 reliability estimate will be an adequate estimate of alternate-form reliability. However, because *Praxis* tests are used to make pass/fail decisions, information about the reliability of classification (RELCLASS) also is relevant to the issue of test reliability. RELCLASS is described in more detail on page 36.

Standard Error of Measurement

The standard error of measurement (SEM) is an estimate of the standard deviation of the distribution of observed scores around a theoretical true score. The SEM can be interpreted as an index of expected variation if the same test taker could be tested repeatedly on different forms of the same test without benefiting from practice or being hampered by fatigue. The SEM of a raw score is computed from the reliability estimate (r_x) and the standard deviation (SD_x) of the scores by the formula:

$$SEM_x = SD_x \sqrt{1 - r_x} .$$

The standard error of measurement for the scaled score is:

$$SEM_s = A * SEM_x .$$

where A is the score conversion coefficient used in the scaled score conversion equation:

$$Scaled\ Score = A * (raw\ score) + B .$$

When the raw-to-scaled score conversion for a test form is nonlinear, the A parameter is estimated using the ratio of the scaled score standard deviation to the raw score standard deviation.

Estimates of the SEM of the scaled score are provided for many of the *Praxis* tests in Appendix B. When sample sizes for a test form are small, several administrations of the form are accumulated to provide a more accurate estimate of the SEM. When several different forms of a test are available for use, the SEM (reported in Appendix B) is averaged across the forms.

The Conditional Standard Error of Measurement (CSEM) is specific to each score level and, therefore, is able to reflect the errors of measurement associated with low-scoring test takers or high-scoring test takers. CSEMs for *Praxis* tests are computed using Lord's (1984) Method IV, and are included in the *Praxis* Test Analysis Reports.

Reliability of Classification

Since *Praxis* tests are intended for certification, assessing the consistency and accuracy of pass/fail decisions is very important. *Praxis* statistical analysts use the Livingston and Lewis method (1995) to estimate decision accuracy and consistency at each cut-score level. *Classification accuracy* is the extent to which the decisions made on the basis of a test would agree with the decisions made from all possible forms of the test (i.e., an estimate of the test taker true score). *Classification consistency* is the extent to which decisions made on the basis of one form of a test would agree with the decisions made on the basis of a parallel, alternate form of the test.

The estimated percentages of test takers correctly (classification accuracy) and consistently classified (classification consistency) tend to increase in value as the absolute value of the standardized difference (SSD) between the mean total score and the qualifying score increases. When the mean score of test takers is well above or below the qualifying score, the number of test takers scoring at or near the qualifying score is relatively small. Therefore, with fewer test takers in the region of the qualifying score, the number of test takers that could easily be misclassified decreases and the decision reliability statistics reflect that fact by increasing in value.

Reliability of Scoring

The reliability of the scoring process for *Praxis* tests that include constructed-response items is determined by a multi-step process.

1. The inter-rater correlations for each item are obtained from the two independent ratings, and the inter-rater reliabilities are computed from them using the Spearman-Brown formula.
2. Variance errors of scoring for each item are calculated by multiplying the item's variance by $(1 - r_{cis})$, where r_{cis} is the item's inter-rater reliability.
3. The variance errors of scoring for all of the items are added together to form the variance of errors of scoring for the entire test.
4. The standard error of scoring is defined as the square root of the variance errors of scoring for the sum obtained in step 3.

Standard errors of scoring are shown in Appendix B for all *Praxis* tests that include CR items. Please note that the standard errors of scoring for SR tests are zero, as the recording of item responses for these tests is performed mechanically, not by human judgment.

Scoring Methodology

Scoring

For tests consisting only of SR items, a raw score is the number of correct answers on the test. There is no penalty imposed for incorrect responses to SR items.

For tests that include both SR and CR questions, raw scores are a weighted composite of the raw SR score and the scores on the individual CR items. A test taker's score in the SR portion of the test is the sum of the number of items answered correctly. The CR section of the test is scored according to the specifications detailed in the Study Companion documents, at <http://www.ets.org/praxis/prepare/materials>.

For each question, the written responses are read and scored by two qualified scorers who are trained to score the responses to that item according to a pre-specified scoring rubric⁷. The ratings that the scorers assign are based on a rubric developed by educators who are specialists in the subject area. All scorers receive training before they score operational responses. The score on any single CR test item is the sum of the scores for CR items as assigned by the two scorers.

Scoring Methodology for Constructed-Response Items

A CR item is one for which the test taker must produce a response, generally in writing. Such items are designed to probe a test taker's depth of understanding of a content area that cannot be assessed solely through SR items. The time suggested for a response can vary from 10 minutes to 60 minutes. Scoring can be:

- Analytic by focusing on specific traits or features
- Holistic by focusing on the response as a whole
- Focused holistic by blending analytic and holistic

Test developers are responsible for the creation of scoring guides, the selection of samples for training purposes, and the training of scoring leadership in test content and scoring standards and procedures.

Every test that contains CR items has a General Scoring Guide (GSG), which is written to verify that well-trained, calibrated scorers will be able to consistently evaluate responses according to clearly specified indicators. Question-specific scoring guides (QSSG) and scoring notes also are developed to inform scorers of some of the item-specific features that a response might contain. Final ratings are assigned to a response after a careful reading to find the evidence that the item has been answered. That evidence then is evaluated by selecting the set of descriptors in the scoring guide that best fits the evidence. This rating can be on various scales, such as 0-3 or 0-6, depending on how much evidence an item is designed to elicit from test takers.

⁷ For many tests, if there is a discrepancy of more than one point between the scores assigned by the two scorers, a third person scores the response. For some tests, "back readings," or third readings, are carried out on a subsample of a certain percentage of responses.

Scoring guides for new items are developed as the prompt is developed and are further refined during sample selection before the first scoring of a prompt. Sample selection is the process during which the chief reader and question leaders for a given test:

- Read through the test takers' responses
- Find responses at each score point on the score scale for the test
- Agree on how to score the selected responses
- Document the rationales for the agreed-upon scores
- Arrange the selected responses into training and calibrating sets for each question on a test

After a scoring guide is finalized during its first use, it can be changed only under very narrowly defined conditions and with approval from the statistical coordinator for the test.

The goals of scoring a response according to a GSG, for a test as well as a QSSG, can be summarized as follows to verify:

- That a test taker receives a fair and appropriate score
- That all test takers are rated in the same manner using the same criteria
- That scoring is conducted consistently throughout a scoring session and from one scoring session to another

To verify the standardization of the scoring process, the following materials must be developed for every CR item:

- Benchmark samples: exemplars of each score point on the score scale, usually at the mid-range of a score point
- Training samples: responses used to train scorers in the variety of responses that can be expected across the range of each of the points of the scoring guide, often presenting unique scoring issues
- Annotations for the responses (evidence sheets): supplemental information used to explain why samples received the given score, providing consistency in what is said during training
- Calibration samples: responses that have been previously scored and are used to assess whether a scorer has learned how to adequately apply the scoring guides to determine a score. Scorers are said to be calibrated when their individual ratings on a set of common CR responses are consistent with scores assigned by other scorers (known also as the "set score"). If a scorer's scores are not consistent with the set score, then she/he is required to be retrained. Calibration verifies to some degree that ratings assigned to a given CR response by different scorers within and between different testing administrations are not very discrepant.

- Training manuals: an outline of the process that a scoring leader should follow in training scorers

In addition, for certain tests Trend Scoring is used as a quality control measure:

- *Praxis*® CR test forms are sometimes used at more than one test administration. At the second and succeeding administrations of the form, the *Praxis* program requires the rescoring of samples of responses from a previous administration of that form (which are seeded into the operational responses) when the test volume is adequate. This procedure is known as Trend Scoring. If the original scores assigned to the trend samples differ on average from the ratings assigned at the rescoring, a shift in scorer severity is presumed. When this occurs, the test form is re-equated using the rescored responses as an external anchor. If there is no change in scorer severity, the raw-score-to-scaled-score conversion used for the earlier administration of the test is retained. (It should be noted that the rescoring of old responses is carried out only to determine if scorer differences have occurred over time, and *not* for the purpose of revising the test scores already reported to test takers at the earlier test administration.)

Scoring leaders are responsible for direct training of scorers as well as overseeing the quality of scoring. Their responsibilities include:

- Assisting in selecting training materials
- Conducting scorer training and, if necessary, retraining
- Monitoring scoring through backreading and counseling scorers
- Verifying that all scoring procedures are followed
- Recommending scorers for scoring leadership
- Scorers are responsible for reading at a sustained rate and giving appropriate scores based on established criteria. They are practicing educators and higher education faculty who are familiar and knowledgeable with the test content.

Consistency in the scoring of a form is verified by:

- Training notes that clearly indicate how an item should be interpreted
- Annotations that clearly indicate how individual responses should be scored as well as the rationale for the score
- Scoring notes that may focus on providing content-related information for scorers
- Training procedures that are outlined and scripted
- Bias training to minimize possible impact of bias that scorers may bring to the scoring session
- Calibration of scorers to ensure that they perform the scoring consistently from administration to administration



Content Category Information

On many *Praxis* tests, items are grouped into content categories. To help test takers in further study or in preparing to retake the test and to help other score users (e.g., the institutions of higher education), the score report shows how many “raw points” have been earned in each content category.

For a test consisting only of SR items, “raw points” means the number of items answered correctly. For tests that include CR items, “raw points” include the sum of the ratings that the scorers awarded to the CR answers as well as the SR raw points. Some SR/CR tests assign scoring weights to the CR section to adjust its contribution to the total raw points available

ETS provides educator preparation programs (EPPs) with the same level of individual student category information that the company provides to test takers because of EPPs’ desire to assist test takers in developing study plans and to have information about the effectiveness of their test takers’ preparation. Although this information is currently being supplied, ETS cautions that category scores are less reliable than total test scores, given the reduced number of items measuring a category. They also may be less reliable because category scores are not equated across forms, so test taker variability in any given category may be due to differences in content difficulty. ETS encourages EPPs to consider other information about a student’s understanding in addition to category scores when making instructional decisions for students.

Quality Assurance Measures

SR answer sheets are machine scored, which gives a high degree of accuracy. However, occasionally test takers feel their scores have been reported incorrectly. In such cases, test takers may request verification of a test score if they feel the score is in error. (Responses to SR items on computer-delivered tests are automatically verified before scores are reported.)

All CR scorers have been carefully trained and follow strict scoring procedures. Most CR items are scored by more than one scorer. However, test takers may still request that their scores be verified for a test that includes CR items if they feel that the score does not accurately reflect their performance. For CR items, this service consists of having a scorer review the responses and the ratings to determine if the ratings are consistent with the scoring rules established for that test.

Appropriate Score Use

ETS is committed to furthering quality and equity in education by providing valid and fair tests, research, and related services. Central to this objective is helping those who use the *Praxis*® tests to understand what are considered their proper uses. The booklet [*Proper Use of The Praxis Series and Related Assessments*](#) defines proper test use as adequate evidence to support the intended use of the test and to support the decisions and outcomes rendered on the basis of test scores.

Proper assessment use is a joint responsibility of ETS as the test developer, and of states, agencies, associations, and institutions of higher education as the test users. The *Praxis* program is responsible for developing valid and fair assessments in accordance with technical guidelines established by the American Educational Research Association, the American Psychological Association, and the



National Council on Educational Measurement in Education ([*Standards for Educational and Psychological Testing*](#), (2014).

Test users are responsible for selecting a test that meets their credentialing or related needs, and for using that test in a manner consistent with the test's intended and validated purpose. Test users must validate the use of a test for purposes other than those intended and supported by existing validity evidence. In other words, they must be able to justify that the intended alternate use is acceptable.

Both ETS and test users share responsibility for minimizing the misuse of assessment information and for discouraging inappropriate assessment use.

Score Reporting

Score reporting is the process in which tests are graded and test results are reported to test takers, institutions, and state agencies.

Scanning/Scoring

ETS has the capacity to score approximately 64,000 tests per day. For SR items, detailed scanning and scoring procedures are done by computer, providing virtually 100 percent accuracy. Established quality-control procedures ensure error-free scanning of all SR answer documents. CR tests utilize group and online scoring sessions that allow ETS to engage practicing educators nationwide and within particular states.

Score Reports

Each test taker receives a detailed score report that includes the test taker's overall score, passing status and, if applicable, information regarding performance on specific areas of the test. The report also includes explanatory materials to help the test taker understand the scoring, such as:

- The scoring process
- Frequently asked questions about scores
- A glossary of important terms used in scoring
- A list of passing scores in the state for all *Praxis*® tests

Following each test administration, depending on state reporting guidelines, scores also are reported to:

- Colleges and universities
- State departments of education
- The American Speech-Language-Hearing Association (ASHA)
- The National Association of School PsychologistsSM (NASPSM)
- Department of Defense Education Activity (DODEA)
- Any other entity designated to receive scores by the state or law.

Score Information for States and Institutions

When score reports are released to the test taker, score information also is released to the applicable state department of education and to those institutions of higher education that the test taker has designated to receive score reports. Score reports contain current scores as well as highest scores earned by the test taker on each test taken in the past ten years. The reports also include basic information on each test taker, such as age, gender, major area of study, GPA, and degree status.

Scores are reported to states, agencies, and institutions through the ETS® Data Manager application.

The Quick and Custom Analytical Reports service within the ETS Data Manager for the *Praxis* tests allows states and IHEs the ability to analyze test taker data. Quick and Custom Analytical Reports offers many analytical functions, including sophisticated searching, data comparison, and chart and table creation.

Users can view data for different test-taker groups based on variables such as gender, ethnicity, educational level and type of educator preparation program. Test-taker groups are customized for the individual user to ensure the privacy of test takers as well as the individual agencies and institutions that serve them

Title II Reporting

Overview

ETS provides a reporting procedure and deliverables, which allow states and institutions to comply with federal reporting requirements on the quality of their teacher preparation programs. These requirements are commonly known as Title II.

In October 1998, Congress voiced concern for the quality of teacher preparation by enacting Title II of the Higher Education Act (HEA). Title II authorizes accountability measures in the form of reporting requirements for institutions and states on teacher preparation and licensing. It is the hope of the U.S. Department of Education, and the desire of Congress, that institutions and states use the reports in meaningful ways to improve teacher education in America.

Section 207 of Title II requires the annual preparation and submission of three reports on teacher preparation and licensing: one from institutions to states, a second from states to the U.S. Secretary of Education, and a third from the Secretary of Education to Congress and the public.

The U.S. Department of Education developed a Reference and Reporting Guide to provide definitions and reporting procedures to help states and institutions supply the information that section 207 requires in timely, uniform, and accurate reports. The implementation procedures that states adopt must be in accordance with state laws and, to the extent possible, reflect existing relationships between institutions and states.

In this three-stage reporting process:

1. Institutions report to their states on several items related to their teacher preparation programs, such as size and composition of their programs.
2. States provide data on its requirements for initial licensure or certification, and compile a more comprehensive report that covers all teacher preparation programs within the state.
3. The Department of Education compiles all state reports into a national report.

By law, these reports must be submitted annually. The Reference and Reporting Guide prescribes the timeframe for reporting, calculation methods, and the data that institutions and states must report.

Submission of the required institutional and state pass rates is a complex process. For example, while institutions of higher education know the names of program completers, they do not



necessarily have complete records of their *Praxis* test scores because students often do not designate their colleges as a score recipient. ETS's Title II services manage the logistical complexities for its clients.

Customized Reporting

To help client states and their teacher preparation programs comply with the congressional mandate, an ETS database stores the specific annual licensure requirements for each state, including licensure tests and passing-score requirements. This ensures that the correct passing score is used in calculating each passing rate. In addition, only tests that are part of the requirements for a student's license are reported.

ETS integrates this database system with a secure Web application to manage enrolled students' data for each teacher preparation program.

This database system:

- Collects enrolled students' data from each teacher preparation program
- Matches each enrolled students' information with the correct test by licensure area
- Lists all enrolled students by their licensure area, test, test category, match status, or update status.

Client Support

Communication is the hallmark of a smooth and successful reporting system. ETS conducts live and recorded webinars to provide states and teacher preparation programs with:

- Information and updates on reporting requirements
- A demonstration of the ETS Title II Web site
- Answers to questions about Title II.

ETS assists each institution with the use of the Web application, and provides information on collecting its enrolled students' data, schedules for relevant due dates, and statistical support in interpreting the passing-rate data. ETS also provides user-friendly formatted reports, both Summary and Single Assessment, in advance of final federal submission so that all institutions have an opportunity to view their data for accuracy.

ETS also maintains a telephone hotline for state users and email service for institutional users to respond to Title II queries. These mechanisms allow ETS to respond to concerns or questions from state agencies or teacher preparation programs.

Appendix A — The *Praxis*® Job Analyses

The following is a list of job analyses related to the *Praxis*® tests. ETS updates its job analysis studies periodically. New studies also are conducted as tests are added to the *Praxis* licensure assessments.

Job Analyses

Agriculture	Middle School Social Studies
Art	Music
Biology	Physical Education
Braille Proficiency	Physics
Business Education	Pre-Kindergarten Education
Chemistry	Principles of Learning and Teaching
Core Academic Skills for Educators: Mathematics	Professional School Counselor
Core Academic Skills for Educators: Reading	Psychology
Core Academic Skills for Educators: Writing	Reading Specialist
Driver Education	School Leader
Early Childhood Education	School Psychologist
Earth and Space Sciences	School Superintendent
Economics	Social Studies
Elementary Education	Sociology
English Language Arts	Special Education: Core Knowledge and Applications
English to Speakers of Other Languages	Special Education: Core Knowledge and Mild to Moderate Applications
Environmental Education	Special Education: Core Knowledge and Severe to Profound Applications
Family and Consumer Sciences	Special Education: Education of Deaf and Hard of Hearing Students
Fundamental Subjects	Special Education: Preschool/Early Childhood
General Science	Special Education: Teaching Speech to Students with Language Impairments
Geography	Special Education: Teaching Students with Behavioral Disorders/Emotional Disturbances
Gifted Education	Special Education: Teaching Students with Intellectual Disabilities
Government/Political Science	Special Education: Teaching Students with Learning Disabilities
Health Education	Special Education: Teaching Students with Visual Impairments
Latin	Speech Communication
Library Media Specialist	Teaching Reading
Marketing Education	Teaching Reading: Elementary Education
Mathematics	Technology Education
Middle School English Language Arts	Theatre
Middle School Mathematics	World and U.S. History
Middle School Science	World Languages

Appendix B – Statistical Characteristics of the *Praxis*® Core Academic Skills for Educators Tests, the *Praxis*® Subject Assessments, and School Leadership Series Tests

Table 1 in this section provides important scoring and statistical information for many of the *Praxis* tests. Notes at the end of the table provide more information about the data included.

- **Range** — The lowest to the highest scaled score possible on any edition of the test. The actual maximum and minimum possible scores for a given form of a test may differ from one edition of a test to another.
- **Interval** — The number of points separating the possible score levels. If the score interval is 10, for example, only scores divisible by 10 are possible.
- **Number of Test Takers** — The number of people taking the test within the time period listed in the notes following the table.
- **Median** — The score that separates the lower half of the scores from the upper half, calculated for the scores obtained by the group of test takers listed in the notes following the table.
- **Average Performance Range** — The range of scores earned by the middle 50 percent of the test takers, calculated for the group of test takers listed in the notes following the table. This range provides an indication of the difficulty of the test.
- **Mean** — The arithmetic average, calculated for the scores obtained by the group of test takers listed in the notes following Table 1.
- **Standard Deviation** — The amount of variability among the scores obtained by the group of test takers listed in the notes following Table 1.
- **Standard Error of Measurement** — The standard error of measurement (SEM) is a test statistic described on page 35 that is often used to characterize the reliability of the scores of a group of test takers. A test taker's score on a single administration of a test will differ somewhat from the score the test taker would receive on another occasion. The more consistent a test taker's scores are from one testing to another, the smaller the SEM. Because estimates of the standard error may vary slightly from one test administration to another and from one test edition to another, the tabled values are averages of the SEMs obtained from all forms of the test currently in use.

- Standard Error of Scoring** — For tests in which the scoring involves human judgment, this statistic describes the reliability of the process of scoring the test takers' responses. A test taker's score on one of these tests will depend to some extent on the particular scorers who rate her/his responses. The more consistent the ratings assigned to the same responses by different scorers, the smaller the standard error of scoring (SES). If a large number of test takers take a test for which the standard error of scoring is four points, about two-thirds of them will receive scores within four points of the scores that they would get if their responses were scored by all possible scorers. The SES is included in Table 1 for tests in the *Praxis*® assessments that include CR items. The tabled values are averages of the SESs obtained from all forms of the test currently in use. Since the January 2008 *Praxis* test administration, all CR tests have been scored by two independent raters. The standard error of scoring for a test consisting only of SR items is zero, because SR scoring is a purely mechanical process with no possibility of disagreement between scorers.
- Reliability** — The reliability coefficient is an estimate of the correlation between test takers' test scores and the scores they might have achieved on different forms of the same test. Its value ranges from zero to one. This index is calculated using an internal consistency estimate (Kuder and Richardson, 1937), based on the statistical relationships among the test takers' responses to all items in the test. The reliability of a test may vary slightly from one test administration to another and from one form of the test to another. The tabled values are averages of the reliabilities obtained from all the forms of the test currently in use.

Table 1 — Statistical Characteristics of *Praxis* Core Academic Skills for Educators Tests, *Praxis* Subject Assessments, and School Leadership Series Tests

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Core Academic Skills for Educators: Mathematics (5732)	100 - 200	2	68779	156	142 – 168	154.9	21.1	8.4	0.0	0.874
Core Academic Skills for Educators: Reading (5712)	100 – 200	2	61935	172	162 – 184	171.4	17.7	7.1	0.0	0.871
Core Academic Skills for Educators: Writing (5722)	100 – 200	2	66628	164	158 – 170	163.7	11.6	5.5	2.1	0.820
Agriculture (5701)	100 – 200	1	940	169	160 – 178	168.4	12.4	5.2	0.0	0.854
Algebra I (5162)	100 – 200	1	806	165	154 – 179	165.9	17.1	7.4	0	0.795
American Sign Language (0634)	100 – 200	1	15	160	140 – 190	157.3	29.5	i	i	i
Art: Content and Analysis (5135)	100 – 200	1	2182	166	158 – 173	164.9	13.4	6.3	2.1	0.814
Art: Content Knowledge (5134)	100 – 200	1	3589	165	158 – 173	164.3	14.2	5.3	0.0	0.833
Audiology (5342)	100 – 200	1	1942	179	174 – 184	178.7	9.0	5.0	0.0	0.807
Biology: Content Knowledge (CT) (5235)	100 – 200	1	8591	163	153 – 175	163.5	15.5	4.3	0.0	0.935
Braille Proficiency (0633)	100 – 200	1	35	178	164 – 186	175.5	14.9	i	i	i



Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Business Education: Content Knowledge (5101)	100 – 200	1	4804	172	162 – 181	170.5	13.9	5.0	0.0	0.896
Chemistry: Content Knowledge (CT) (5245)	100 – 200	1	3635	160	149 – 174	160.4	19.5	5.6	0.0	0.925
Chinese (Mandarin): World Language (5665)	100 – 200	1	339	195	185 – 199	187.0	19.0	4.2	1.4	0.941
Citizenship Education: Content Knowledge (5087)	100 – 200	1	84	168	153 – 177	166.4	15.0	i	i	i
Computer Science (5652)	100 – 200	1	f	f	f	f	f	f	f	f
Connecticut Administrator Test (6412)	100 – 200	1	f	f	f	f	f	f	f	f
Early Childhood: Content Knowledge (5025)	100 – 200	1	7239	171	159 – 180	168.0	16.6	5.6	0.0	0.895
Earth and Space Sciences: Content Knowledge (5571)	100 – 200	1	1810	165.5	155 – 178	164.7	17.3	5.2	0.0	0.914
Economics (5911)	100 – 200	1	375	152	140 – 164	153.2	18.7	6.2	0.0	0.910
Education of Young Children (5024)	100 – 200	1	4971	169	162 – 177	168.0	13.7	5.3	1.4	0.844
Educational Leadership: Administration and Supervision (5411)	100 – 200	1	9408	165	156 – 173	164.5	12.2	5.8	0.0	0.799
Educational Leadership: Administration and Supervision (5412)	100 – 200	1	f	f	f	f	f	f	f	f
Elementary Education: Applied Mathematics—CKT (7903)	100 – 200	1	f	f	f	f	f	f	f	f
Elementary Education: Applied Reading and Language Arts—CKT (7902)	100 – 200	1	f	f	f	f	f	f	f	f
Elementary Education: Content Knowledge (5018)	100 – 200	1	16665	169	160 – 179	167.3	17.5	5.7	0.0	0.892
Elementary Education: Curriculum, Instruction, and Assessment (5017)	100 – 200	1	9694	171	162 – 179	170.4	12.5	5.9	0.0	0.832
Elementary Education: Instructional Practice and Applications (5019)	100 – 200	1	5417	169	160 – 178	168.2	14.4	5.4	1.9	0.875



Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Elementary Education: Mathematics—CKT (7803)	100 – 200	1	38	163.5	150 – 172	159.9	19.3	i	i	i
Elementary Education: Mathematics Subtest (5003)	100 – 200	1	38982	172	161 – 186	171.3	19.9	9.4	0.0	0.807
Elementary Education: Reading and Language Arts—CKT (7802)	100 – 200	1	38	169	156 – 179	167.0	16.4	i	i	i
Elementary Education: Reading and Language Arts Subtest (5002)	100 – 200	1	37647	170	161 – 179	169.3	14.2	7.4	0.0	0.777
Elementary Education: Science—CKT (7804)	100 – 200	1	40	182	162 – 188	174.9	18.4	i	i	i
Elementary Education: Science—CKT (7904)	100 – 200	1	f	f	f	f	f	f	f	f
Elementary Education: Science Subtest (5005)	100 – 200	1	37899	168	161 – 179	168.2	16.0	8	0.0	0.787
Elementary Education: Social Studies Subtest (5004)	100 – 200	1	37920	166	156 – 177	165.9	16.5	7.9	0.0	0.791
Elementary Education: Social Studies (7805)	100 – 200	1	40	158.5	142 – 184	162.5	21.8	i	i	i
Elementary Education: Social Studies (7905)	100 – 200	1	f	f	f	f	f	f	f	f
English Language Arts: Content and Analysis (5039)	100 – 200	1	6859	175	169 – 181	173.9	10.8	4.7	2.2	0.823
English Language Arts: Content Knowledge (5038)	100 – 200	1	14825	179	171 – 186	177.5	12.5	4.6	0.0	0.879
English to Speakers of Other Languages (5362)	100 – 200	1	6151	177	168 – 185	176.2	12.7	5.0	0.0	0.849
Environmental Education (0831)	100 – 200	1	19	176	169 – 190	177.4	12.3	4.9	0.0	0.882
Family and Consumer Sciences (5122)	100 – 200	1	2495	163	156 – 170	162.4	10.8	4.8	0.0	0.836
French: Content Knowledge (5174)	100 – 200	1	973	172	160 – 184	170.6	18.1	5.1	2.5	0.923
Fundamental Subjects: Content Knowledge (5511)	100 – 200	1	4085	175	164 – 185	173.6	15.1	5.8	0.0	0.876
General Science: Content Knowledge (5435)	100 – 200	1	6824	164	152 – 178	164.2	18.7	5.4	0.0	0.922



Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Geography (5921)	100 – 200	1	272	173	162 – 181	171.3	15.5	5.1	0.0	0.898
German: World Language (5183)	100 – 200	1	293	176	161 – 194	173.6	22.2	5.1	2.5	0.953
Gifted Education (5358)	100 - 200	1	1397	164	158 – 170	163.9	10.0	5.4	0.0	0.731
Government/Political Science (5931)	100 - 200	1	707	168	156 – 179	166.2	16.2	5.4	0.0	0.913
Health and Physical Education: Content Knowledge (5857)	100 – 200	1	5116	165	160 – 173	164.2	12.6	5.6	0.0	0.813
Health Education (5551)	100 – 200	1	3065	166	157 – 174	164.7	12.3	5.4	0.0	0.785
Interdisciplinary Early Childhood Education (5023)	100 - 200	1	532	181	174 – 186	179.6	8.3	4.6	0.0	0.749
Kentucky Specialty Test of Instructional and Administrative Practices (6015)	100 – 200	1	1121	170	164 – 179	172.0	10.9	4.9	0.0	0.825
Latin (5601)	100 – 200	1	130	181.5	164 – 195	178.9	17.5	5.1	0.0	0.951
Library Media Specialist (5311)	100 - 200	1	3390	164	156 – 171	163.5	11.7	4.4	0.0	0.880
Marketing Education (5561)	100 - 200	1	631	171	159 – 178	167.8	15.8	5.8	0.0	0.869
Mathematics: Content Knowledge (5161)	100 - 200	1	15974	159	137 – 169	154.1	22.5	7.3	0.0	0.875
Middle School English Language Arts (5047)	100 – 200	1	7039	165	154 – 172	162.7	13.2	5.7	2.3	0.783
Middle School Mathematics (5169)	100 - 200	1	13252	170	157 – 180	167.1	18.4	6.8	0.0	0.854
Middle School Science (5440)	100 - 200	1	6900	159	146 – 171	157.3	19.3	6.3	0.0	0.902
Middle School Social Studies (5089)	100 - 200	1	5141	166	154 – 179	165.7	18.0	6.3	2.2	0.915
Middle School: Content Knowledge (5146)	100 - 200	1	3888	162	151 – 174	162.3	16.8	6.2	0.0	0.884
Music: Content and Instruction (5114)	100 - 200	1	2577	167	160 – 174	165.2	13.4	6.3	1.8	0.757
Music: Content Knowledge (5113)	100 - 200	1	5138	168	160 – 176	167.5	12.9	5.8	0.0	0.832
ParaPro Assessment (1755)	420 - 480	1	78863	470	462 – 476	467.7	10.42	3.4	0.0	0.941
Pennsylvania Grades 4-8 Core Assessment: English Language Arts and Social Studies (5154)	100 – 200	1	3869	162	152 – 174	162.8	16.4	8.3	0.0	0.783
Pennsylvania Grades 4-8 Core Assessment: Mathematics and Science (5155)	100 – 200	1	3868	171	160 – 183	169.9	18.9	8.3	0.0	0.814



Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Pennsylvania Grades 4-8 Core Assessment: Pedagogy (5153)	100 – 200	1	2642	179	172 – 185	178.7	9.9	5.9	0.0	0.685
Pennsylvania Grades 4-8 Subject Concentration: English Language Arts (5156)	100 – 200	1	1371	168	158 – 180	167.3	17.5	7.0	0.0	0.849
Pennsylvania Grades 4-8 Subject Concentration: Mathematics (5158)	100 – 200	1	1610	174	156 – 184	169.6	21.0	8.1	0.0	0.830
Pennsylvania Grades 4-8 Subject Concentration: Science (5159)	100 – 200	1	809	161	152 – 173	161.6	16.9	6.9	0.0	0.838
Pennsylvania Grades 4-8 Subject Concentration: Social Studies (5157)	100 – 200	1	568	162	150 – 176	162.3	18.8	7.4	0.0	0.849
Physical Education: Content and Design (5095)	100 - 200	1	2730	170	162 – 175	167.2	12.4	5.4	2.0	0.783
Physical Education: Content Knowledge (5091)	100 - 200	1	4545	155	150 – 160	154.3	8.9	3.9	0.0	0.817
Physics: Content Knowledge (CT) (5265)	100 - 200	1	2003	152	138 – 168	151.7	22.5	5.9	0.0	0.934
Pre-Kindergarten Education (5531)	100 – 200	1	223	180	169 – 186	176.9	12.9	5.7	0.0	0.800
Principles of Learning and Teaching: Grades 5-9 (5623)	100 - 200	1	5470	176	168 – 182	174.7	10.6	5.4	2.3	0.786
Principles of Learning and Teaching: Grades 7-12 (5624)	100 - 200	1	31630	176	169 – 183	175.2	11.1	5.7	2.5	0.813
Principles of Learning and Teaching: Early Childhood (5621)	100 - 200	1	6631	169	162 - 176	168.5	11.0	5.4	2.3	0.779
Principles of Learning and Teaching: Grades K-6 (5622)	100 - 200	1	30084	177	169 – 183	175.6	10.3	5.1	2.1	0.812
Professional School Counselor (5421)	100 - 200	1	9257	170	163 – 177	169.2	9.9	4.4	0.0	0.858
Psychology (5391)	100 - 200	1	297	170	159 – 178	169.2	14.5	5.5	0.0	0.876
Reading for Virginia Educators: Elementary and Special Education (5306)	100 - 200	1	7441	176	166 – 184	174.6	13.5	5.6	1.6	0.885



Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Reading for Virginia Educators: Reading Specialist (5304)	100 - 200	1	733	183	175 – 191	181.5	13.8	6.6	1.4	0.871
Reading Specialist (5301)	100 - 200	1	4151	183	175 - 190	181.9	11.9	5.9	1.8	0.791
School Leaders Licensure Assessment (6011)	100 - 200	1	18000	176	169 – 181	174.6	9.1	4.6	1.8	0.772
School Leaders Licensure Assessment (6990)	100 - 200	1	f	f	f	f	f	f	f	f
School Psychologist (5402)	100 - 200	1	6354	170	162 – 177	169.0	10.8	4.5	0.0	0.852
School Superintendents Assessment (6021)	100 - 200	1	1886	169	164 – 175	169.1	8.4	4.6	1.8	0.718
Social Studies: Content and Interpretation (5086)	100 - 200	1	3831	161	150 – 171	159.6	16.6	5.6	2.1	0.874
Social Studies: Content Knowledge (5081)	100 - 200	1	12819	167	157 – 178	167.0	15.1	4.6	0.0	0.919
Sociology (5952)	100 – 200	1	105	172	164 – 181	171.9	11.4	i	i	i
Spanish: World Language (5195)	100 - 200	1	5522	175	160 – 186	170.9	20.1	5.4	2.3	0.936
Special Education: Core Knowledge and Mild to Moderate Applications (5543)	100 - 200	1	11840	172	164 – 179	171.1	11.6	4.6	2.0	0.858
Special Education: Core Knowledge and Severe to Profound Applications (5545)	100 - 200	1	1827	177	171 – 183	176.4	9.7	4.1	1.8	0.839
Special Education: Teaching Speech to Students with Language Impairments (5881)	100 - 200	1	72	164.5	156 – 173	164.7	12.2	i	i	i
Special Education: Core Knowledge and Applications (5354)	100 - 200	1	14097	174	165 – 181	172.5	12.2	5.1	0.0	0.847
Special Education of Deaf and Hard of Hearing Students (5272)	100 - 200	1	366	169	164 – 176	168.8	9.9	5.3	0.0	0.759
Special Education: Preschool/Early Childhood (5691)	100 – 200	1	1726	175	168 – 181	174.2	9.7	4.4	0.0	0.836
Special Education: Teaching Students with Behavioral Disorders/Emotional Disturbances (5372)	100 - 200	1	1282	179	171 – 185	177.0	11.6	4.9	0.0	0.826



Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Special Education: Teaching Students with Intellectual Disabilities (5322)	100 - 200	1	171	179	172 – 184	177.1	11.5	5.1	0.0	0.806
Special Education: Teaching Students with Learning Disabilities (5383)	100 - 200	1	504	170	160 – 177	167.2	14.0	5.4	0.0	0.880
Special Education: Teaching Students with Visual Impairments (5282)	100 - 200	1	367	169	163 – 176	168.4	11.9	5.5	0.0	0.832
Speech Communication: Content Knowledge (5221)	100 – 200	1	704	160	151 – 169	159.6	13.0	4.7	0.0	0.868
Speech-Language Pathology (5331)	100 – 200	1	21766	178	171 – 184	177.0	10.5	5.0	0.0	0.861
Teaching Reading (5204)	100 - 200	1	4898	168	161 – 177	168.3	11.2	4.7	1.6	0.878
Teaching Reading: Elementary Education (5203)	100 – 200	1	9636	176	169 – 182	175.4	10.3	4.6	1.7	0.854
Technology Education (5051)	100 – 200	1	1703	180	170 – 189	178.6	14.0	5.1	0.0	0.880
Theatre (5641)	100 – 200	1	825	170	161 – 179	168.9	14.2	5.1	0.0	0.862
World and U.S. History: Content Knowledge (5941)	100 – 200	1	2677	162	150 – 172	161.0	16.6	5.2	0.0	0.919
World Languages: Pedagogy (5841)	100 - 200	1	594	181	172-190	179.4	14.2	6.8	1.9	0.787

Notes:

- “Number of Test Takers,” “Median,” and “Average Performance Range” were calculated from the records of test takers who took the test between August 2015 and July 2018, and who are in the particular educational group described below. If a test taker took the test more than once in this period, the most recent score was used. Test takers were selected according to their responses to the question, “What is the highest educational level you have reached?” These statistics are provided if the test was taken by 30 or more test takers in the specified time period.
- The Median and Average Performance Range for the Core Academic Skills for Educators tests were calculated on college freshmen, sophomores, and juniors.
- The Median and Average Performance Range for all other tests were calculated on test takers who were college seniors, college graduates, graduate students, or holders of master’s or doctoral degrees.

Legend:

c = Constructed-response items were consensus-scored.

i = Insufficient data

f = New test. Data not yet available.

Bibliography

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.

Dorans, N., & Holland, P. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. Holland and H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355–368.

Educational Testing Service, *ETS Guidelines for Fairness Review of Assessments*, Princeton, N.J., 2009.

Educational Testing Service, *ETS Standards for Quality and Fairness*, Princeton, N.J., 2014

Educational Testing Service, *Questions to Ask About Teacher Testing*, Princeton, N.J., 2004

Educational Testing Service, *Proper Use of the PRAXIS Assessments and Related Assessments*, Princeton, N.J., 2016

Holland, P.W. & Thayer, D.T. (1985). An alternative definition of the ETS delta scale of item difficulty (RR-85-43). Princeton, N.J.: Educational Testing Service.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), *Test Validity*, pp. 129–145. Hillsdale, N J: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Joint Committee on Testing Practices (2005). *Code of Fair Testing Practices in Education*. Washington, D.C.

Knapp, J., & Knapp, L. (1995). Practice analysis: Building the foundation for validity. In J.C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 93–116). Lincoln, NE: Buros Institute of Mental Measurements.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices* (2nd Ed.). New York: Springer-Verlag.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.

Livingston, S.A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.

Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, 21, 239–243.

Raymond, M.R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education*, 14, 369–415.

Schmitt, K (1995). What is licensure? In J.C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 3–32). Lincoln, NE: Buros Institute of Mental Measurements.

Tannenbaum, R.J., & Rosenfeld, M. (1994). Job analysis for teacher competency testing: Identification of basic skills important for all entry-level teachers. *Educational and Psychological Measurement*, 54, 199–211

Von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The kernel method of equating. New York: Springer.

Wainer, H. & Kiely, G. (1987). Item clusters and computer adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185–201.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement*, 30, 233–251.



www.ets.org