ETS PRAXIS

# Technical Manual for the *Praxis*® Tests

## and Related Assessments

ETS® Professional Educator PROGRAMS

October 2021

# Table of Contents

# 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 use of *Praxis* test scores
- 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 *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. *Praxis* tests provide states with testing tools and ancillary services that support their teacher licensure and certification process. These tools include tests of academic skills and subject-specific assessments related to teaching.

All states want teachers to have the knowledge and skills needed for safe and effective practice before they receive a license. To address this desire, *Praxis* tests are designed to assess test takers' job-relevant knowledge and skills. States adopt *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 *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 the *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 that was influenced by 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 to maintain multiple test forms, which is often the case with smaller, single-state testing programs.

States also customize their selection of the *Praxis* assessments. *Praxis* frequently has 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 needs. States also customize their passing-score requirements on *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

The *Praxis* Core Academic Skills for Educators (or *Praxis* Core) 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 educator preparation 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.

States that have chosen to use one or more of the *Praxis* Subject Assessments require their applicants to take the tests as part of the teacher licensure process. Each *Praxis* test is designed to provide states with a standardized way to assess whether prospective teachers have demonstrated knowledge that is 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 domains for the *Praxis* Subject Assessments are 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 educators 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.) Teachers and faculty who prepare teachers in the content area are involved in multistate standard-setting studies to recommend passing (or cut) scores to state agencies responsible for educator licensure.

## 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 state 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. 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. Information to help decision makers understand the critical issues associated with teacher assessment programs is available on the States and Agencies portion of the *Praxis* website.

In addition, ETS has developed a guide, *Proper Use of the* Praxis *Series and Related Assessments* (PDF) 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 many helpful test preparation tools**.** These materials include:

- 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, such as "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. Institutions also can use the ETS Data Manager to produce annual summary reports of their *Praxis* test takers' scores. The *Praxis* Program also offers 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 providing tests 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 standards[1].

[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 licensure[2] 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 measured by the

[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."

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, *Praxis* Subject Assessments, and the School Leadership Series 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 advisory committees, 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 *Praxis* educator licensure tests rely on educators throughout the design and development process to ensure that the tests are valid for their intended purpose. Practicing educators and college faculty who prepare educator candidates are involved from the definition of the content domains through the design of test blueprints and development of test content.

The content tested on *Praxis* Subject tests is fundamentally based on available national and state standards for the field being assessed. The development process begins with a committee of educators who use the national standards to draft knowledge and skill statements that apply to beginning educators. This ***Development Advisory Committee (DAC)*** is facilitated by an experienced ETS assessment specialist. The draft knowledge and skill statements created by this group are then presented via an online survey to a large sample of educators who are asked to judge (a) the relevance and importance of each statement for beginning practice and (b) the depth of knowledge that would be expected of a beginning educator. This ***Job Analysis Survey*** also gathers relative importance (i.e., weights) for the categories within the draft content domain.

A second committee of educators, the ***National Advisory Committee (NAC)***, is convened to review the draft content domain and the results of the Job Analysis Survey to (a) further refine the content domain for the test, (b) develop the test specifications or blueprint, and (c) determine the types of test questions that will be used to gather evidence from test takers. The resulting test specifications are then presented in a second online survey by a large sample of educators to confirm that the content

[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.

of the test includes knowledge and skills relevant and important (i.e., weights) for beginning practice. The results of the ***Confirmatory Survey*** are used by the NAC and ETS assessment specialists to finalize the 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 (PDF).

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

Both the Development Advisory Committee and the National Advisory Committee are assembled to be diverse with respect to

- race, ethnicity, and gender,
- practice settings, grade levels, and geographic regions, and
- professional perspectives.

Such diversity and representation reinforce the development of the content domain knowledge and/or skills that is applicable across the profession and supports the develop of tests that are considered fair and reasonable to all test takers.

### *Validity Maintenance*

ETS assessment specialists work closely with educators on an ongoing basis to monitor national associations and other relevant indicators to determine whether revisions to standards or other events in the field may warrant changes to a licensure test. ETS also regularly gathers information from educator preparation programs and state licensure agencies to assure that the tests are current and meeting the needs of the profession. If significant changes have occurred, the process described above is triggered. Routinely, ETS conducts an online ***Test Specification Review Survey*** to determine whether the test continues to measure relevant and important knowledge and skills for beginning educators. Gathering validity evidence is not a single event but an ongoing process.

## Test Development Process

Following the development of test specifications (described above), *Praxis* tests and related materials follow a rigorous development process, as outlined below and in Figure 1:

- 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.
- Conduct virtual and in-person meetings with educators to fulfill the development of the test specifications for the specific content.
- Develop enough 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 (PDF) (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.

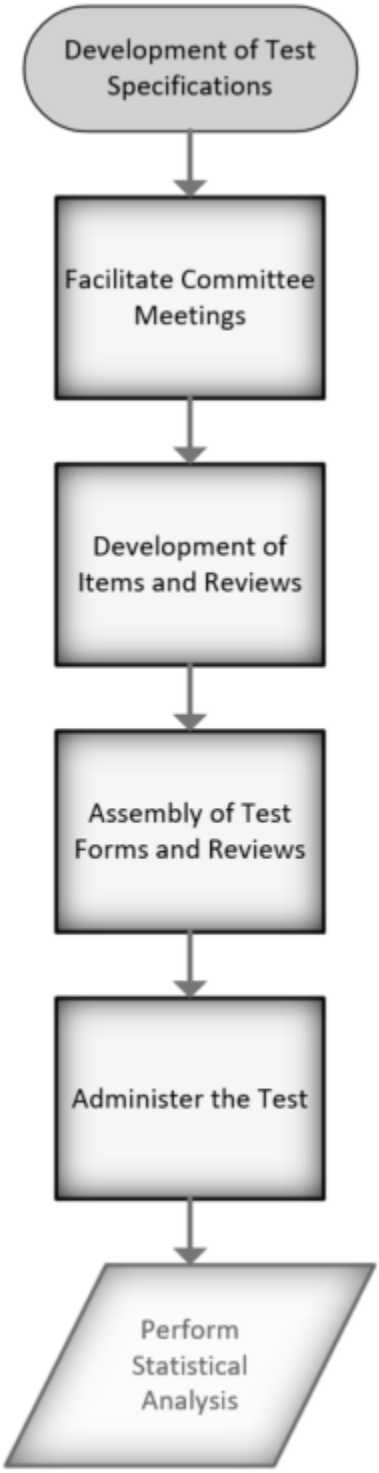


**Figure 1:** Test Development Process

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

### *Development of Test Specifications*

The test specifications are developed jointly between ETS test developers and external educators with the specific content knowledge for the area being developed.

### *Facilitate Committee Meetings*

Educators are recruited from *Praxis* user states to participate in virtual and in-person meetings to provide input into the depth and breadth of the knowledge and skills needed for a beginning teacher. These educators range from novice teachers (1-7 years) in the content area to the more veteran teachers and well as the educator preparation program professors.

### *Development of Test Items and Reviews*

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 enough evidence that the test taker is competent to begin practice.

The outside item development 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.

Items received are then sent through an extensive content review process with internal ETS test developers, fairness reviewers, and editors. Resolution of the items are completed along the review path and are documented. The final content review and sign-off of the items is completed prior to the item being ready for use on a form.

### *Assembly of Test Forms and Review*

ETS test developers assemble a test form(s) using items that have been reviewed and approved by content experts, fairness, and edit. A preview of the items selected to be used in a form is then generated for test developers to check for quality. 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. If any changes are made to the items, they 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.

When content reviews of a 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 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.

### *Perform Statistical Analysis*

Once enough responses have been gathered, 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 to score an item, a Problem Item Notice (PIN) is issued and distributed. The distribution of a PIN triggers actions in the Psychometric Analysis & Research, 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 enough 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. 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 ETS Standards for Quality and Fairness (PDF). These standards reflect a commitment to producing fair, valid, and reliable tests and are applied to all ETS-administered programs. Compliance with the standards 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 (PDF).

In addition to complying with the ETS quality standards, ETS tests comply with the Standards for Educational and Psychological Testing (2014) and The Code of Fair Testing Practices in Education (PDF).

### *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*

**Educator Licensure**. The *Praxis* Core Academic Skills for Educators tests may be used by the licensing body or agency within a state for teacher licensing 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 Educator Preparation Programs**. These tests also may be used by institutions of higher education to identify students with enough reading, writing, and mathematics skills to enter a 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 scores as one criterion for permitting students to move on to the student teaching phase of their program. 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 (e.g., 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:<br>• Accepts the *Praxis* Subject Assessment as valid<br>• Sets a schedule for meeting content-based licensure requirements<br>• 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 that students were exposed to covers that content.

The following summarizes this process:

| IF… | THEN… |
|---|---|
| an institution, but not the state, requires content-based licensure before student teaching is allowed<br><br>AND<br><br>the state requires the use of a *Praxis* Subject Assessment content test for licensure | 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 |

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 prerequisite for entrance into student teaching and the curriculum which students were exposed to covers 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 *Praxis* Subject Assessments in a state that has not authorized the content assessment for licensure<br><br>AND<br><br>the state requires use of a *Praxis* content test 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. |

**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 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 *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 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 12-18 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 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. Two rounds of judgments are collected, with panelist discussion during the second round. A panelist's judgments are summed across SR items to calculate that panelist's individual passing score; the mean of the panelists' passing scores is reported and the recommended passing score of the panel.

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 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 scores, as defined by the rubric. Multiple rounds of judgments are collected, with panelist discussion during the second round. As with the method used for SR items, a panelist's judgments are summed across CR items to calculate that panelist's individual passing score; the mean of the panelists' passing scores is reported and the recommended passing score of the panel.

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, participating and interested states receive 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' Psychometric Analysis & Research division developed procedures designed to support the calculation 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 (PDF) (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

When a new selected-response form is introduced, a Preliminary Item Analysis (PIA) of the test items is completed before other analyses are conducted . 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 typically combined to yield a total score for each test question.

A Differential Item Functioning (DIF) Analysis is conducted to verify 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 are 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 once sufficient test taker responses have been acquired.

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 flagged 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 enough 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 conducted once sufficient data has been acquired. All standard test takers who have a raw total score and answer at least three selected-response items in a test form are included in the item analyses.

Preliminary and final item 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.

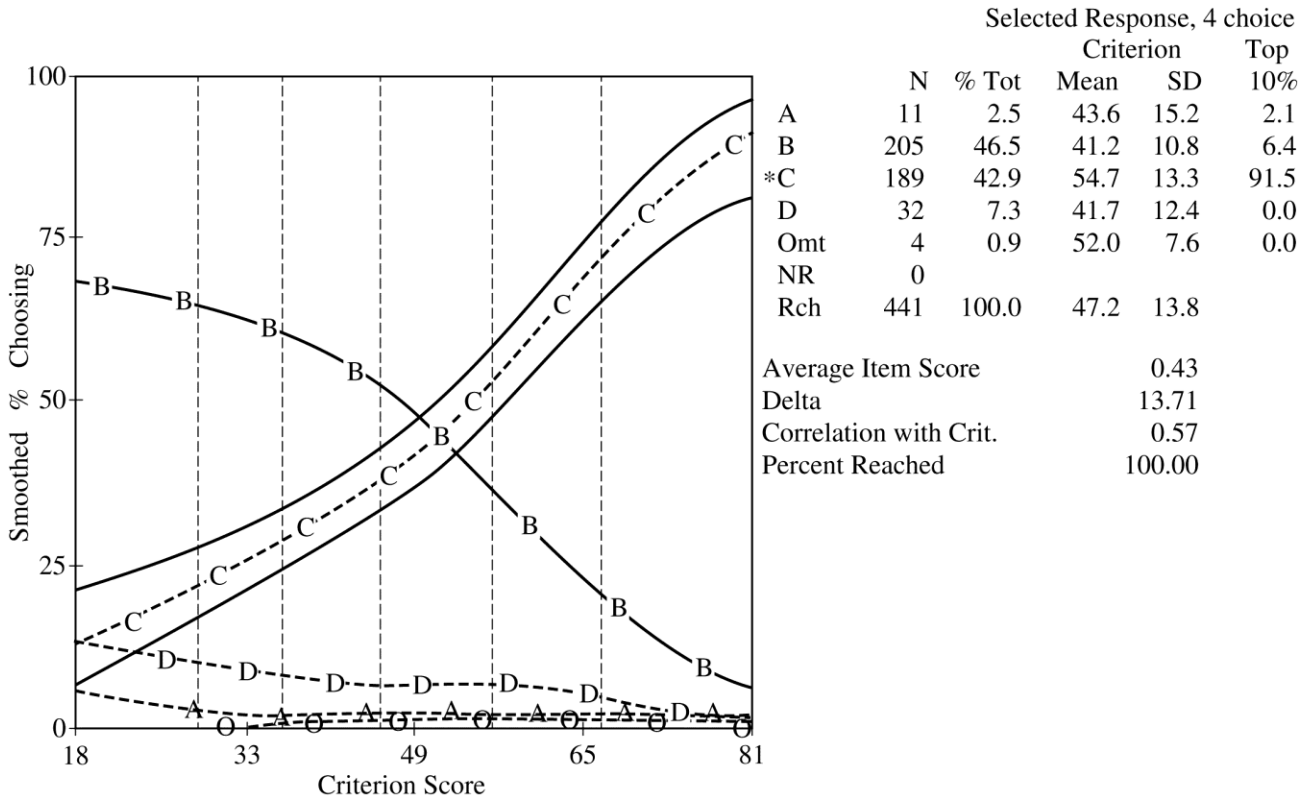


**Figure 2.** Example of an item analysis graph for an SR item

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 10th, 25th, 50th, 75th, and 90th 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.

After statistical analysts review a PIA, they deliver the result 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 SiGNET 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 takers[4]
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 test takers reached all or nearly all items, it would be wrong to conclude, without additional information, that the test (or

[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.

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 once sufficient test taker responses have been acquired. For example, DIF analysis can be used to measure the fairness of test items at a test taker subgroup level. Only standard test takers who answer at least three selected-response items and indicate that English is their best language of communication and that they first learned English or English and another language as a child are included in DIF analyses. 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, 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 enough volume to do so.

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 enough 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
- 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 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 – 4*z*, 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
  - absolute value less than 1.5, or
  - 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 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 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* (PDF) 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 = Test Length * (1 / 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: 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 = (Scale Maximum – Scale Minimum) / (T – C), and

   B = (Scale Maximum – Scale Minimum) – (A * C)

### *Equating*

To maintain the comparability of the reported scores derived from different forms of the same test, the forms are equated. First the initial form of the test is scaled. Each subsequent form of the test is administered and then placed on the scale through the equating process prior to score reporting. The equating transforms the raw scores of each form to the adjusted scaled scores that can be reported. The equating procedures consider the difficulty of the form and the relative ability of the group of test takers who took that form. All standard test takers who have a raw total score and answer at least three selected-response items in a test form are included in the equating sample.

### *The NEAT Design*

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* Psychometric Analysis & Research 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.

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 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 when a new form is introduced, 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 many 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 SiGNET Design*

The current equating practices explained above are not appropriate for very low volume tests (e.g., those tests that have fewer than 100 test takers consistently per administration). For these tests, the *Praxis* program uses the single-group equating with nearly equivalent test forms (SiGNET) design. In this design, the test is constructed of a number of item clusters (also 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 SiGNET test is conducted to establish a raw-to-scaled score conversion for its first administrations. When enough 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 SiGNET 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 enough 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.

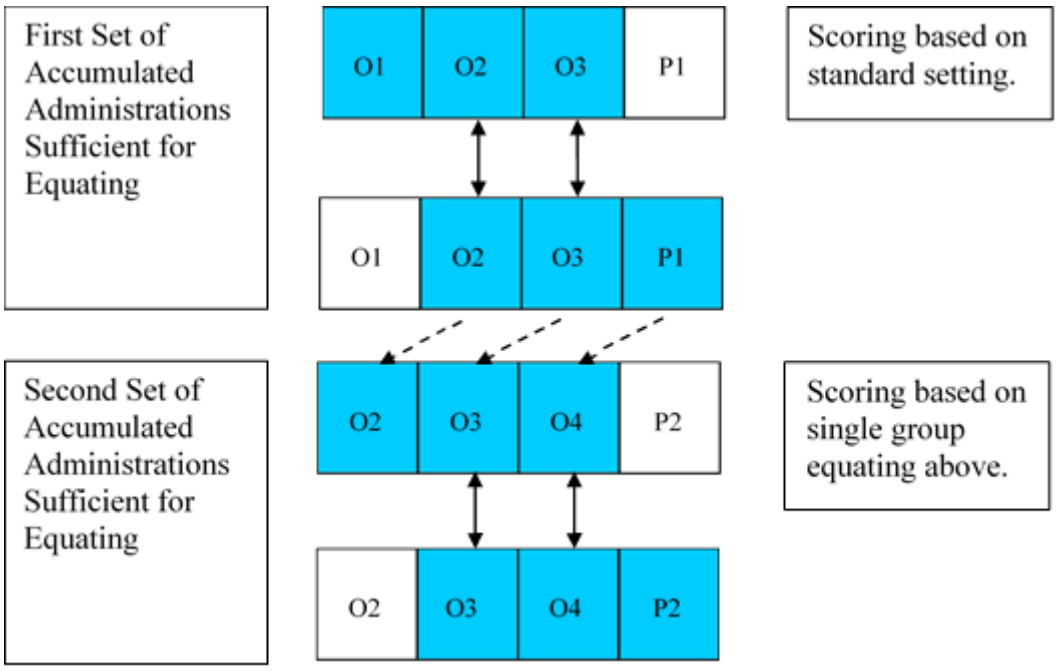


**Figure 3.** The SiGNET Equating Design

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

### *The ISD Design*

For computer-based tests without constructed-response items and with moderate to high volumes to move to continuous testing, the newly conceptualized Interchangeable Section Design (ISD) with item response theory (IRT) pre-equating is implemented. With this design, tests are broken into sections, called testlets, either according to content domains, with each testlet containing one or more content categories, or as mini-tests, with each testlet mimicking the full test. Multiple versions of each testlet are created, which are considered interchangeable with the same content specification and statistical characteristics. Each test title contains operational and pretest testlets. During test delivery, the system randomly selects a version from each of the operational and pretest testlets and creates a test form on the fly according to the test specifications. This way, an exponential number of form combinations can be generated to reduce security concerns and to accumulate data for IRT calibration. See **Figure 4** for an illustration.

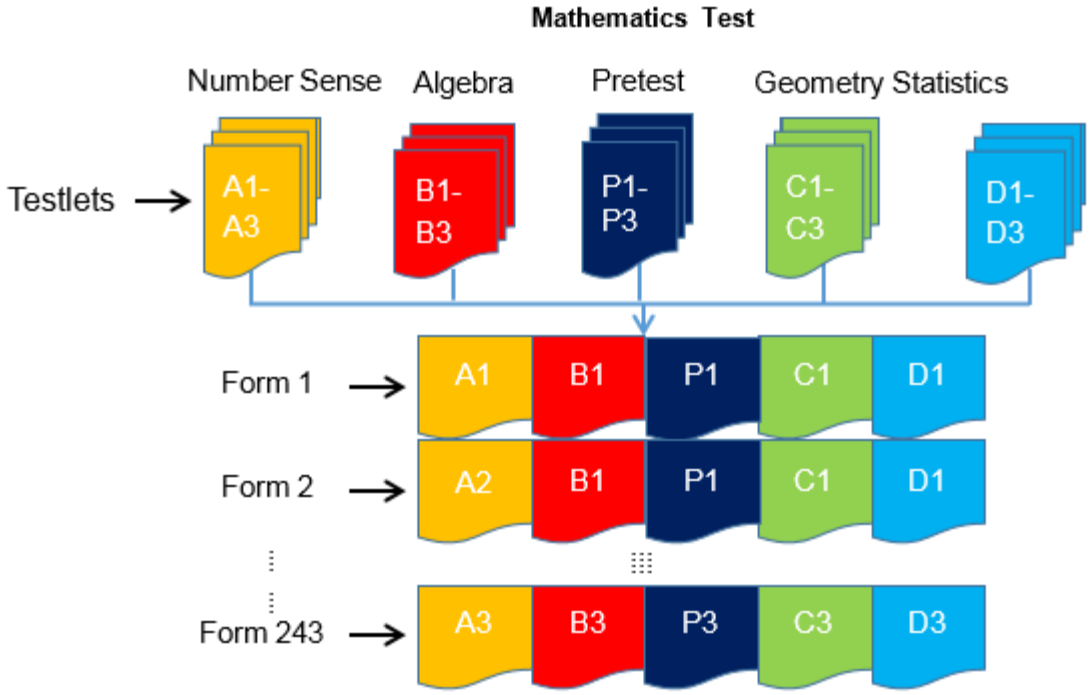


**Figure 4.** An Example of ISD Design

Implement the ISD design requires accumulation of data on existing forms to establish an adequate item pool to assemble the necessary testlets and to conduct pre-equating. In the pre-equating, IRT models (i.e., one-parameter logistic model, or two-parameter logistic model) are used to calibrate pools of items. Once items are calibrated and estimates of their difficulties, discriminations, and other parameters are obtained, new testlets are assembled and raw score-to-scaled score conversions are constructed. The conversion table is obtained by converting the ability estimate based on test taker's raw score (e.g., the number of correct responses) to a scaled score, with a common transformation relationship to all combinations of the testlets.

*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 SiGNET design. All other *Praxis* tests use the NEAT design to equate new test forms. The detailed equating methodologies (e.g., equating design, rational for selection, and equating methods, etc.) for the *Praxis* Core Academic Skills for Educators Tests, the *Praxis* Subject Assessments, and the School Leadership Series Assessments are summarized in Table 1.

**Table 1**. Equating Methodology Summary

| Assessments | Equating Design | Selection Rationale* | Equating Methods |
|---|---|---|---|
| The *Praxis*® Core Academic Skills for Educators Tests | The Equivalent Groups Design | Extremely high-volume tests ( >1000 per admin per form) | Set Means and SDs Equal; Direct Equipercentile |
| | The NEAT Design | During the first several administrations, testing volumes are not stable, but tests still have moderate to high volumes ( >100 per admin) | Tucker, Levine, Chained Linear, Chained Equipercentile |
| The *Praxis* Subject Assessments | The NEAT Design | Moderate to high-volume tests ( >100 per admin) | Tucker, Levine, Chained Linear, Chained Equipercentile |
| | The SiGNET Design | Low-volume tests (<100 per admin consistently) | Set Means and SDs Equal; Direct Equipercentile |
| | ISD-IRT | Moderate to high-volume selected-response-only tests moving to continuous testing | IRT concurrent calibration and true score equating |
| The School Leadership Series Assessments | The NEAT Design | Moderate to high-volume tests ( >100 per admin) | Tucker, Levine, Chained Linear, Chained Equipercentile |
| | The SiGNET Design | Low-volume tests (<100 per admin consistently) | Set Means and SDs Equal; Direct Equipercentile |

* The selection rationale for a certain equating design depends on a number of factors: the testing volume, the test format, the score reporting and delivery schedule, and some other factors such as defective items, out-of-date content, unwanted exposure, etc.

## 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 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:

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

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 A. 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 A) is averaged across the forms.

The Conditional Standard Error of Measurement (CSEM) is specific to each score level and, therefore, can 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 based on 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 based on one form of a test would agree with the decisions made based on 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 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 A 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.

For most *Praxis*® tests, the written responses on each CR question are read and scored by two qualified human scorers who are trained to score the responses to that item according to a pre-specified scoring rubric[7]. 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.

Automated machine scoring for CR questions is implemented in several *Praxis* tests, namely, the e-rater® for the *Praxis* Core Academic Skills for Educators Writing Test, and the c-rater™ model for several *Praxis* subject assessments on reading. Both e-rater and c-rater are automated ETS scoring engines developed based on data from thousands of previously scored essays. The e-rater is a generic scoring model that scores both argumentative prompts and source-based prompts, while the c-rater scores responses to content-based, short-answer questions. For the above tests, each CR response are scored by a human scorer and the e-rater or the c-rater. If the human score and the automated machine score agree, the two scores are added to become the final score for the CR question. If they differ by more than a specified amount, the CR response is rated by a different human scorer, whose rating is used to resolve the discrepancy.

## 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

[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.

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.

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
- 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

Parallel scoring methods are used to score spoken responses, e.g., in World Languages tests.

## Content Category Information

On *Praxis* tests, items are classified by the content category they assess. 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* (PDF) 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.

## Scoring

ETS delivers over 40 million scores and 250 million score reports annually with an on-time score release rate of 99.999 percent. Scoring for constructed-response 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 Psychologists℠ (NASP℠)
- 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 – Statistical Characteristics of the Praxis® Core Academic Skills for Educators Tests, the Praxis® Subject Assessments, and School Leadership Series Tests

The table 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 2.

**Standard Deviation** — The amount of variability among the scores obtained by the group of test takers listed in the notes following Table 2.

**Standard Error of Measurement** — The standard error of measurement (SEM) is a test statistic described on page 36 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 2 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 2 — 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 |
|---|---|---|---|---|---|---|---|---|---|---|
| Agriculture (5701) | 100 - 200 | 1 | 969 | 168 | 159 - 176 | 167.7 | 12.2 | 5.2 | 0 | 0.85 |
| Algebra I (5162) | 100 - 200 | 1 | 1262 | 166 | 153 – 178 | 164.1 | 19.1 | 7.2 | 0 | 0.82 |
| American Sign Language (0634) | 100 - 200 | 1 | 30 | 160 | 150 – 170 | 156.7 | 18.1 | i | i | I |
| Art: Content and Analysis (5135) | 100 - 200 | 1 | 1990 | 165 | 157 – 172 | 163.7 | 13.5 | 5.8 | 2.3 | 0.80 |
| Art: Content Knowledge (5134) | 100 - 200 | 1 | 3522 | 165 | 158 – 173 | 164.3 | 13.9 | 5.5 | 0 | 0.84 |
| Audiology (5342) | 100 - 200 | 1 | 1549 | 178 | 173 – 184 | 178.1 | 9.0 | 5 | 0 | 0.82 |
| Audiology (5343) | 100 - 200 | 1 | 505 | 167 | 162 – 173 | 166.8 | 11.2 | 5.6 | 0 | 0.81 |
| Biology: Content Knowledge (CT) (5235) | 100 - 200 | 1 | 6948 | 163 | 153 - 174 | 162.8 | 15.4 | 4.3 | 0 | 0.93 |
| Braille Proficiency (0633) | 100 - 200 | 1 | 30 | 179 | 167 - 196 | 179.2 | 19.0 | i | i | I |
| Business Education: Content Knowledge (5101) | 100 - 200 | 1 | 4354 | 171 | 163 - 180 | 170.3 | 13.7 | 4.9 | 0 | 0.90 |
| Chemistry: Content Knowledge (5245) | 100 - 200 | 1 | 2742 | 160 | 149 - 173 | 159.8 | 19.3 | 5.6 | 0 | 0.92 |
| Chinese (Mandarin): World Language (5665) | 100 - 200 | 1 | 324 | 196 | 190 - 199 | 190.3 | 16.1 | 4.2 | 1.4 | 0.94 |
| Citizenship Education: Content Knowledge (5087) | 100 - 200 | 1 | 87 | 166 | 155 - 178 | 165.9 | 15.3 | 5.2 | 0 | 0.90 |
| Communication and Literacy: Reading (5714) | 100 - 200 | 2 | f | F | f | f | f | f | f | f |
| Communication and Literacy: Writing (5724) | 100 - 200 | 2 | f | f | f | f | f | f | f | f |
| Computer Science (5652) | 100 - 200 | 1 | 823 | 165 | 149 - 186 | 165.1 | 23.3 | 6.1 | 0 | 0.94 |
| Connecticut Administrator Test (6412) | 100 - 200 | 1 | 1224 | 172 | 164 - 178 | 170.4 | 9.9 | 5.7 | 0 | 0.717 |
| Core Academic Skills for Educators: Mathematics (5733) | 100 - 200 | 2 | 21806 | 168 | 154 - 182 | 166.2 | 21.5 | 7.7 | 0 | 0.89 |
| Core Academic Skills for Educators: Reading (5713) | 100 - 200 | 2 | 18976 | 170 | 158 - 184 | 169.7 | 18.6 | 7.5 | 0 | 0.87 |
| Core Academic Skills for Educators: Writing (5723) | 100 - 200 | 2 | 21477 | 164 | 154 - 170 | 161.9 | 13.0 | 6.3 | 1.9 | 0.80 |
| Early Childhood Education (5025) | 100 - 200 | 1 | 8414 | 171 | 160 - 181 | 168.8 | 16.4 | 5.5 | 0 | 0.90 |
| Early Childhood: Mathematics (5028) | 100 – 200 | 1 | f | f | f | f | f | f | f | f |

| Test | Scale Range | Interval | No. of Test Takers | Median | Average Performance Range | Mean | Standard Deviation | Standard Error of Measurement | Standard Error of Scoring | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| Early Childhood: Reading and Language Arts and Social Studies (5027) | 100 – 200 | 1 | f | f | f | f | f | f | f | f |
| Earth and Space Sciences: Content Knowledge (5571) | 100 - 200 | 1 | 1532 | 163 | 152 - 175 | 162.1 | 17.8 | 5.2 | 0 | 0.91 |
| Economics (5911) | 100 - 200 | 1 | 293 | 155 | 141 - 167 | 155.4 | 20.8 | 6.2 | 0 | 0.91 |
| Education of Young Children (5024) | 100 - 200 | 1 | 5188 | 169 | 161 - 177 | 167.5 | 13.7 | 5.4 | 1.9 | 0.84 |
| Educational Leadership: Administration and Supervision (5412) | 100 - 200 | 1 | 7352 | 168 | 159 - 175 | 166.4 | 11.3 | 5.7 | 0 | 0.78 |
| Elementary Education: Content Knowledge (5018) | 100 - 200 | 1 | 7370 | 168 | 158 - 178 | 166.1 | 17.4 | 5.7 | 0 | 0.89 |
| Elementary Education: Curriculum, Instruction, and Assessment (5017) | 100 - 200 | 1 | 5823 | 170 | 161 - 178 | 169.1 | 13.2 | 5.9 | 0 | 0.83 |
| Elementary Education: Mathematics and Science (5008) | 100 – 200 | 1 | f | f | f | f | f | f | f | f |
| Elementary Education: Mathematics—CKT (7813) | 100 - 200 | 1 | 8125 | 160 | 150 - 172 | 158.8 | 18.4 | 7.625 | 0 | 0.82 |
| Elementary Education: Mathematics Subtest (5003) | 100 - 200 | 1 | 42021 | 173 | 161 - 186 | 171.1 | 20.3 | 9.2 | 0 | 0.82 |
| Elementary Education: Three Subject Bundle—Mathematics (5903) | 100 - 200 | 1 | 856 | 164 | 146 - 179 | 160.8 | 23.5 | 9.2 | 0 | 0.82 |
| Elementary Education: Reading and Language Arts—CKT (7812) | 100 - 200 | 1 | 3310 | 170 | 161 - 178 | 169.1 | 13.8 | 6.775 | 0 | 0.76 |
| Elementary Education: Reading and Language Arts Subtest (5002) | 100 - 200 | 1 | 41972 | 170 | 161 - 179 | 168.9 | 14.1 | 6.8 | 0 | 0.80 |
| Elementary Education: Reading and Language Arts & Social Studies (5007) | 100 – 200 | 1 | f | f | f | f | f | f | f | f |
| Elementary Education: Science—CKT (7814) | 100 - 200 | 1 | 3241 | 175 | 165 - 186 | 173.6 | 17.2 | 7.975 | 0 | 0.80 |
| Elementary Education: Science Subtest (5005) | 100 - 200 | 1 | 42299 | 169 | 161 - 179 | 168.8 | 15.9 | 7.9 | 0 | 0.78 |

| 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: Three Subject Bundle—Science (5905) | 100 - 200 | 1 | 875 | 164 | 152 - 174 | 161.4 | 18.1 | 7.9 | 0 | 0.78 |
| Elementary Education: Social Studies—CKT (7815) | 100 - 200 | 1 | 3406 | 172 | 164 - 182 | 171.0 | 14.5 | 7.33 | 0 | 0.74 |
| Elementary Education: Social Studies Subtest (5004) | 100 - 200 | 1 | 42508 | 165 | 157 - 177 | 166.4 | 16.3 | 7.9 | 0 | 0.78 |
| Elementary Ed: Three Subject Bundle-Social Studies (5904) | 100 - 200 | 1 | 858 | 160 | 148 - 171 | 159.4 | 17.2 | 7.9 | 0 | 0.78 |
| English Language Arts: Content and Analysis (5039) | 100 - 200 | 1 | 5919 | 174 | 168 - 181 | 172.8 | 11.5 | 4.7 | 2.2 | 0.82 |
| English Language Arts: Content Knowledge (5038) | 100 - 200 | 1 | 13387 | 179 | 171 - 186 | 177.5 | 12.5 | 4.7 | 0 | 0.88 |
| English to Speakers of Other Languages (5362) | 100 - 200 | 1 | 11989 | 177 | 168 - 185 | 176.0 | 12.8 | 5.2 | 0 | 0.85 |
| Environmental Education (0831) | 100 - 200 | 1 | 9 | i | i | i | i | i | i | i |
| Family and Consumer Sciences (5122) | 100 - 200 | 1 | 2492 | 163 | 156 - 170 | 162.5 | 10.7 | 4.8 | 0 | 0.84 |
| French: World Language (5174) | 100 - 200 | 1 | 803 | 172 | 159 - 183 | 169.5 | 18.9 | 5.2 | 2.4 | 0.92 |
| Fundamental Subjects: Content Knowledge (5511) | 100 - 200 | 1 | 4488 | 173 | 161 - 184 | 172.1 | 15.9 | 5.6 | 0 | 0.88 |
| General Science: Content Knowledge (5435) | 100 - 200 | 1 | 5397 | 161 | 150 - 177 | 162.5 | 19.0 | 5.4 | 0 | 0.92 |
| Geography (5921) | 100 - 200 | 1 | 157 | 175 | 164 - 187 | 172.8 | 17.1 | 5.1 | 0 | 0.90 |
| Geometry (5163) | 100 - 200 | 1 | 51 | 152 | 132 - 165 | 150.5 | 21.7 | i | i | i |
| German: World Language (5183) | 100 - 200 | 1 | 238 | 180 | 162 - 193 | 175.3 | 21.5 | 5.2 | 2.2 | 0.96 |
| Gifted Education (5358) | 100 - 200 | 1 | 1646 | 164 | 159 - 170 | 163.8 | 10.0 | 5 | 0 | 0.74 |
| Government/ Political Science (5931) | 100 - 200 | 1 | 487 | 168 | 157 - 179 | 167.0 | 16.4 | 5.5 | 0 | 0.91 |
| Health and Physical Education: Content Knowledge (5857) | 100 - 200 | 1 | 5394 | 165 | 158 - 172 | 163.6 | 12.6 | 5.6 | 0 | 0.81 |
| Health Education (5551) | 100 - 200 | 1 | 2356 | 166 | 156 - 174 | 164.2 | 13.2 | 5.1 | 0 | 0.86 |
| Interdisciplinary Early Childhood Education (5023) | 100 - 200 | 1 | 695 | 181 | 174 - 186 | 179.2 | 9.3 | 4.6 | 0 | 0.75 |
| Japanese: World Language (5661) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |

| Test | Scale Range | Interval | No. of Test Takers | Median | Average Performance Range | Mean | Standard Deviation | Standard Error of Measurement | Standard Error of Scoring | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| Journalism (5224) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |
| Latin (5601) | 100 - 200 | 1 | 103 | 180 | 160 - 199 | 176.9 | 21.6 | 5.1 | 0 | 0.95 |
| Library Media Specialist (5311) | 100 - 200 | 1 | 3142 | 164 | 157 - 172 | 164.0 | 11.7 | 4.5 | 0 | 0.89 |
| Marketing Education (5561) | 100 - 200 | 1 | 522 | 170 | 156 - 178 | 166.3 | 16.4 | 5.8 | 0 | 0.87 |
| Mathematics (5165) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |
| Mathematics: Content Knowledge (5161) | 100 - 200 | 1 | 12763 | 157 | 136 - 169 | 153.1 | 22.8 | 7.2 | 0 | 0.87 |
| Middle School English Language Arts (5047) | 100 - 200 | 1 | 6254 | 164 | 153 - 171 | 162.0 | 13.4 | 5.7 | 2.3 | 0.78 |
| Middle School Mathematics (5164) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |
| Middle School Mathematics (5169) | 100 - 200 | 1 | 11709 | 170 | 158 - 180 | 167.5 | 18.3 | 6.9 | 0 | 0.85 |
| Middle School Science (5440) | 100 - 200 | 1 | 5994 | 159 | 146 - 172 | 157.9 | 19.6 | 6.3 | 0 | 0.90 |
| Middle School Science (5442) | 100 - 200 | 1 | 359 | 156 | 141 - 176 | 157.4 | 23.8 | 5.9 | 0 | 0.94 |
| Middle School Social Studies (5089) | 100 - 200 | 1 | 4526 | 167 | 155 - 180 | 166.0 | 19.2 | 6.2 | 2.3 | 0.88 |
| Middle School: Content Know. (5146) | 100 - 200 | 1 | 788 | 159 | 148 - 172 | 159.0 | 17.8 | 6.2 | 0 | 0.88 |
| Music: Instrumental and General Knowledge (5115) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |
| Music: Vocal and General Knowledge (5116) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |
| Music: Content and Instruction (5114) | 100 - 200 | 1 | 2798 | 166 | 157 - 173 | 164.1 | 13.3 | 6 | 1.7 | 0.77 |
| Music: Content Knowledge (5113) | 100 - 200 | 1 | 4706 | 167 | 160 - 176 | 166.9 | 12.8 | 5.7 | 0 | 0.84 |
| ParaPro Assessment (1755) | 420 - 480 | 1 | 71943 | 470 | 462 - 476 | 467.43 | 10.61 | 3.4 | 0 | 0.94 |
| Pennsylvania Grades 4-8 Core Assessment: English Language Arts and Social Studies (5154) | 100 - 200 | 1 | 3602 | 162 | 152 - 173 | 161.8 | 16.9 | 8.1 | 0 | 0.80 |
| Pennsylvania Grades 4-8 Core Assessment: Mathematics and Science (5155) | 100 - 200 | 1 | 3636 | 171 | 160 - 183 | 169.3 | 19.1 | 8.1 | 0 | 0.82 |
| Pennsylvania Grades 4-8 Core Assessment: Pedagogy (5153) | 100 - 200 | 1 | 1823 | 179 | 172 - 186 | 178.2 | 10.5 | 5.6 | 0 | 0.79 |
| Pennsylvania Grades 4-8 Subject Concentration: English Language Arts (5156) | 100 - 200 | 1 | 972 | 168 | 156 - 179 | 165.9 | 18.2 | 6.9 | 0 | 0.87 |
| Pennsylvania Grades 4-8 Subject Concentration: Mathematics (5158) | 100 - 200 | 1 | 1240 | 175 | 158 - 184 | 169.9 | 20.4 | 7.6 | 0 | 0.84 |

| 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 Subject Concentration: Science (5159) | 100 - 200 | 1 | 669 | 162 | 150 - 173 | 160.8 | 17.7 | 6.9 | 0 | 0.84 |
| Pennsylvania Grades 4-8 Subject Concentration: Social Studies (5157) | 100 - 200 | 1 | 475 | 161 | 150 - 178 | 161.9 | 19.5 | 7.4 | 0 | 0.85 |
| Physical Education: Content and Design (5095) | 100 - 200 | 1 | 2789 | 170 | 161 - 174 | 166.0 | 12.8 | 5.5 | 2.3 | 0.78 |
| Physical Education: Content Knowledge (5091) | 100 - 200 | 1 | 3975 | 155 | 150 - 159 | 153.9 | 8.5 | 3.9 | 0 | 0.83 |
| Physical Science (5485) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |
| Physics: Content Knowledge (5265) | 100 - 200 | 1 | 1763 | 153 | 138 - 167 | 151.5 | 21.1 | 5.9 | 0 | 0.93 |
| Pre-Kindergarten Education (5531) | 100 - 200 | 1 | 177 | 175 | 161 - 181 | 170.8 | 15.8 | 5.7 | 0 | 0.80 |
| Principles of Learning and Teaching: Grades 5-9 (5623) | 100 - 200 | 1 | 4964 | 176 | 168 - 182 | 174.8 | 10.6 | 5.5 | 2.3 | 0.79 |
| Principles of Learning and Teaching: Grades 7-12 (5624) | 100 - 200 | 1 | 23841 | 176 | 168 - 183 | 175.1 | 11.6 | 5.4 | 2.3 | 0.83 |
| Principles of Learning and Teaching: Early Childh'd (5621) | 100 - 200 | 1 | 5128 | 169 | 161 - 176 | 167.9 | 11.6 | 5.4 | 2.2 | 0.78 |
| Principles of Learning and Teaching: Grades K- 6 (5622) | 100 - 200 | 1 | 23753 | 176 | 169 - 183 | 175.2 | 10.3 | 5.1 | 2.1 | 0.81 |
| Principles of Learning and Teaching: Grades PreK-12 (5625) | 100 - 200 | 1 | f | f | f | f | f | f | f | f |
| Professional School Counselor (5421) | 100 - 200 | 1 | 9707 | 170 | 163 - 177 | 169.5 | 9.9 | 4.4 | 0 | 0.86 |
| Psychology (5391) | 100 - 200 | 1 | 260 | 169 | 159 - 182 | 169.9 | 14.5 | 5.1 | 0 | 0.90 |
| Reading for Virginia Educators: Elementary and Special Education (5306) | 100 - 200 | 1 | 6838 | 175 | 165 - 184 | 173.7 | 14.1 | 5.6 | 1.7 | 0.87 |
| Reading for Virginia | 100 - 200 | 1 | 766 | 185 | 176 - 193 | 183.3 | 13.7 | 5.9 | 1.5 | 0.86 |

| Test | Scale Range | Interval | No. of Test Takers | Median | Average Performance Range | Mean | Standard Deviation | Standard Error of Measurement | Standard Error of Scoring | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| Educators: Reading Specialist (5304) | | | | | | | | | | |
| Reading Specialist (5301) | 100 - 200 | 1 | 3681 | 182 | 174 - 189 | 180.6 | 12.2 | 6.1 | 1.8 | 0.78 |
| Reading Specialist (5302) | 100 - 200 | 1 | 97 | 167 | 160 - 175 | 167.4 | 12.7 | i | i | i |
| School Leaders Licensure Assessment (6990) | 100 - 200 | 1 | 10986 | 171 | 162 - 178 | 169.6 | 11.5 | 5.16 | 1.9 | 0.829 |
| School Psychologist (5402) | 100 - 200 | 1 | 7337 | 168 | 161 - 175 | 167.6 | 10.5 | 4.5 | 0 | 0.85 |
| School Superintendents Assessment (6991) | 100 - 200 | 1 | 1044 | 173 | 164 - 180 | 170.6 | 13.4 | c | c | c |
| Social Studies: Content and Interpretation (5086) | 100 - 200 | 1 | 3446 | 160 | 150 - 170 | 158.7 | 17.0 | 5.7 | 2.1 | 0.87 |
| Social Studies: Content Knowledge (5081) | 100 - 200 | 1 | 11082 | 167 | 157 - 178 | 167.1 | 15.4 | 4.6 | 0 | 0.92 |
| Sociology (5952) | 100 - 200 | 1 | 115 | 174 | 166 - 180 | 172.0 | 12.0 | 5.4 | 0 | 0.83 |
| Spanish: World Language (5195) | 100 - 200 | 1 | 4640 | 176 | 162 - 187 | 172.1 | 19.5 | 5.3 | 2.3 | 0.94 |
| Special Education: Core Knowledge and Mild to Moderate Applications (5543) | 100 - 200 | 1 | 12872 | 172 | 165 - 179 | 171.4 | 11.0 | 4.6 | 2 | 0.86 |
| Special Education: Core Knowledge & Severe/Profound Applications (5545) | 100 - 200 | 1 | 1800 | 177 | 170 - 183 | 176.0 | 10.0 | 4.1 | 1.8 | 0.84 |
| Special Education: Teaching Speech to Students with Language Impairments (5881) | 100 - 200 | 1 | 58 | 158.5 | 152 - 165 | 159.1 | 11.1 | 5.5 | 0 | 0.82 |
| Special Education: Core Knowledge and Applications (5354) | 100 - 200 | 1 | 14842 | 173 | 164 - 180 | 171.9 | 12.6 | 4.9 | 0 | 0.86 |
| Special Education of Deaf and Hard of Hearing Students (5272) | 100 - 200 | 1 | 385 | 168 | 161 - 174 | 166.3 | 11.3 | 5.5 | 0 | 0.81 |
| Special Education: Preschool/Early Childhood (5691) | 100 - 200 | 1 | 1791 | 174 | 167 - 180 | 173.3 | 9.9 | 4.4 | 0 | 0.84 |
| Special Education: Teaching Students with Behavioral Disorders/ Emotional | 100 - 200 | 1 | 362 | 174 | 166 - 183 | 173.3 | 12.1 | 4.7 | 0 | 0.87 |

| Test | Scale Range | Interval | No. of Test Takers | Median | Average Performance Range | Mean | Standard Deviation | Standard Error of Measurement | Standard Error of Scoring | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| Disturbances (5372) | | | | | | | | | | |
| Special Education: Teaching Students with Intellectual Disabilities (5322) | 100 - 200 | 1 | 174 | 180 | 172 - 186 | 176.9 | 13.0 | 4.9 | 0 | 0.86 |
| Special Education: Teaching Students with Learning Disabilities (5383) | 100 - 200 | 1 | 545 | 168 | 159 - 177 | 167.0 | 13.5 | 5.2 | 0 | 0.88 |
| Special Education: Teaching Students with Visual Impairments (5282) | 100 - 200 | 1 | 357 | 171 | 164 - 176 | 169.5 | 10.5 | 5.5 | 0 | 0.83 |
| Speech Communication: Content Knowledge (5221) | 100 - 200 | 1 | 512 | 160 | 152 - 169 | 159.6 | 13.1 | 4.7 | 0 | 0.87 |
| Speech-Language Pathology (5331) | 100 - 200 | 1 | 23291 | 177 | 170 - 183 | 176.1 | 10.4 | 5 | 0 | 0.85 |
| Teaching Reading: K-12 (5206) | 100 - 200 | 1 | 758 | 164 | 157 - 172 | 163.7 | 12.1 | 5.2 | 1.9 | 0.82 |
| Teaching Reading: Elementary (5205) | 100 - 200 | 1 | 6893 | 166 | 159 - 174 | 165.4 | 12.6 | 5.1 | 1.5 | 0.83 |
| Technology Education (5051) | 100 - 200 | 1 | 1751 | 180 | 169 - 189 | 177.9 | 14.0 | 5.1 | 0 | 0.88 |
| Theatre (5641) | 100 - 200 | 1 | 824 | 170 | 162 - 179 | 169.4 | 13.1 | 5.2 | 0 | 0.86 |
| World and U.S. History: Content Knowledge (5941) | 100 - 200 | 1 | 2580 | 161 | 150 - 172 | 160.6 | 16.6 | 5.1 | 0 | 0.92 |
| World Languages: Pedagogy (5841) | 100 - 200 | 1 | 502 | 180 | 169 - 189 | 177.5 | 14.8 | 6.8 | 1.9 | 0.79 |

Notes:

"Number of Test Takers," "Median," and "Average Performance Range" were calculated from the records of test takers who took the test between August 2018 and July 2021, 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 Functionin*g. 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.