

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

September 2026

ETS®
**Professional
Educator**
PROGRAMS

Table of Contents

Preface.....	5
Purpose of This Manual.....	5
Audience	5
Purpose of the Praxis Assessments	6
Overview	6
The <i>Praxis</i> Core Academic Skills for Educators Tests	7
The <i>Praxis</i> Subject Assessments — Subject Knowledge and Pedagogical Knowledge Related to Teaching....	7
The School Leadership Series Assessments	7
How the <i>Praxis</i> Assessments Address States’ Needs.....	8
Assessment Development.....	9
Fairness in Test Development	9
Test Development Standards	9
Validity.....	10
Test Development Process.....	13
Review Processes	16
Test Adoption Process	17
Process Overview	17
Analysis of States’ Needs.....	20
Standard-Setting Studies.....	20
Psychometric Methods and Procedures	22
Introduction	22
Item Analyses	22
Classical Item Analyses	22
Item Response Theory Analyses.....	24
Differential Item Functioning Analyses.....	24
Test-Form Equating.....	25
Scaling.....	25
Equating	26
Test Scoring.....	28
Test Statistics.....	29

Reliability.....	29
Standard Error of Measurement.....	29
Reliability of Classification	29
Reliability of Scoring.....	30
Speededness	30
Additional Feedback.....	30
Scoring Methodology	31
Scoring	31
Scoring Methodology for Constructed-Response Items	31
Content Category Information	33
Quality Assurance Measures	34
Appropriate Score Use.....	34
Score Reporting.....	35
Scoring	35
Score Reports	35
Score Information for States and Institutions	35
Title II Reporting.....	36
Appendix A – Statistical Characteristics of the <i>Praxis</i> ® Core Academic Skills for Educators Tests, the <i>Praxis</i> ® Subject Assessments, and School Leadership Series Tests	38
Bibliography	47

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 standards¹.

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

Validity

The Nature of Validity Evidence

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

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

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

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

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

The link forged between occupational content and test content is based on expert judgment by practitioners and other stakeholders in the profession who may have an informed perspective about requisite occupational knowledge and/or skills. Processes for gathering and analyzing content-related validity evidence to support the relevance and importance of knowledge and/or skills 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 development 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 various reviews 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 Guidelines for Developing Fair Tests and Communications (2022) PDF 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.

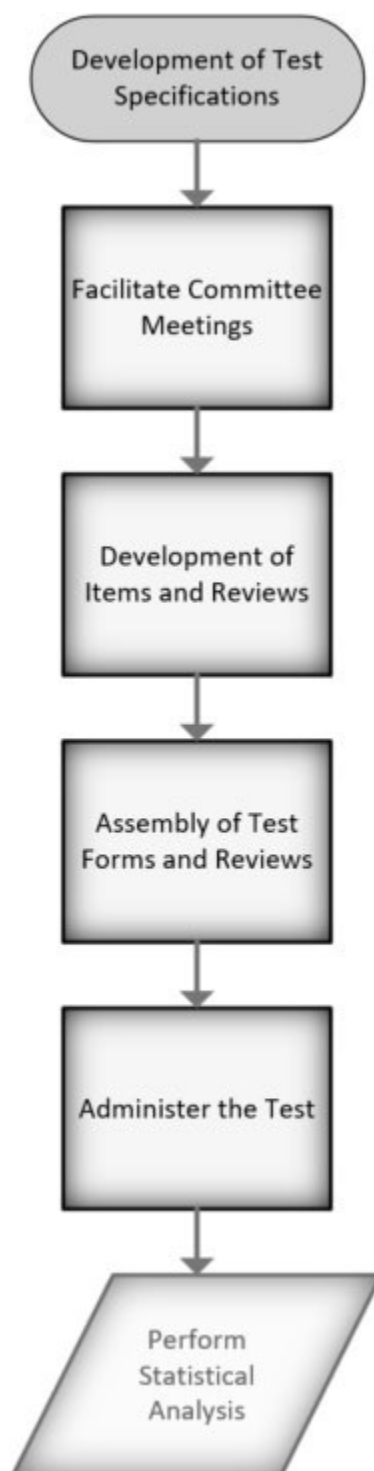


Figure 1: Test Development Process

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 24, 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 Guidelines for Developing Fair Tests and Communications (2022) 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 Guidelines for Developing Fair Tests and Communications (2022) PDF 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 licensure decisions. The *Praxis* program suggests that before adopting a test, the licensing body or agency reviews the test specifications to confirm that the content covered on the test is consistent with state standards and with expectations of what the state's teachers should know and be able to do. The licensing body or agency also must establish a passing standard or "cut score." ETS conducted a multistate standard-setting study for the *Praxis* Core and provided the results to the licensing body or agency to inform its decision.

Entrance into 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: • Accepts the <i>Praxis</i> Subject Assessment as valid • Sets a schedule for meeting content-based licensure requirements • Establishes passing scores to meet requirements

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

AND

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

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

If an institution of higher education wants to use the *Praxis* Subject Assessments but is in a state that has not adopted the identified subject test for teacher licensure, that institution should review the test specifications to confirm that the content covered on the test is a 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 <i>Praxis</i> Subject Assessments in a state that has not authorized the content assessment for licensure	that institution should review test specifications to confirm that the content is necessary for student teaching and that students were exposed to the curriculum that covers the appropriate content.

AND

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

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

However, if a graduate-level program is in a state that has not adopted the subject test, that program should review the test specifications to confirm that the content is a 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 educators of the subject area or college faculty who prepare educators in the subject area and (b) familiarity with the knowledge and skills required of beginning educators. ETS selects panelists to represent the diversity (race/ethnicity, gender, geographic

setting, etc.) of the educator population. Each panel includes approximately 12-18 educators, the majority of whom are practicing, licensed educators 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. Each panelist's judgments are summed across SR items to calculate the 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 Methods and Procedures

Introduction

ETS employs a range of psychometric methods and procedures to support the development, scoring, and reporting of valid, reliable, fair, and comparable scores for the *Praxis*® program. These methods draw on well-established principles of educational measurement and are applied throughout test development, administration, scoring, and score reporting.

The psychometric procedures described in this chapter are consistent with the *Standards for Educational and Psychological Testing* and the *ETS Standards for Quality and Fairness*. These procedures are widely accepted within the psychometric community and support the intended interpretations and uses of *Praxis* scores.

Following the administration of a new test form, statistical analyses are conducted to evaluate item performance, identify potential fairness concerns, support score comparability across test forms, and assess the reliability of reported scores. These analyses include classical item analyses, item response theory (IRT) analyses where applicable, differential item functioning (DIF) analyses, scaling, equating, and other measures of score quality.

The remainder of this chapter describes the methods used to evaluate item performance, monitor fairness, support score comparability across test forms, and assess the quality of reported scores. These procedures provide the psychometric foundation for the interpretation and use of *Praxis* scores.

Item Analyses

Classical Item Analyses

Classical item analyses are conducted to evaluate the statistical performance of test items and to identify items that may require further review. These analyses examine item difficulty, item-total correlations, and response patterns to assess whether items function as intended. Items may be flagged due to unusually high or low difficulty, weak or negative relationships with overall test performance, potential multiple keys, or possible keying errors.

Test developers review flagged items in consultation with content experts or content advisory committees to confirm content accuracy, alignment with the intended construct, and the presence of a single best answer. When necessary, problematic items are removed from scoring and may be revised before future operational use.

Classical Item Difficulty Indices

Item difficulty indices describe the relative difficulty of an item for a group of test takers. Items scored as correct or incorrect are referred to as dichotomous items, whereas items that can receive multiple score points are referred to as polytomous items.

For dichotomous items, difficulty is measured using the p -value, defined as the proportion of test takers who answer the item correctly. P -values range from 0.00 to 1.00, with higher values indicating easier items and lower values indicating more difficult items.

For polytomous items, difficulty is evaluated using the average item score (AIS) or a p -value expressed as the proportion of the maximum possible score earned. The AIS can range from zero to the maximum score available for the item. To facilitate interpretation across item types, polytomous item scores are often expressed as a proportion of the maximum possible score, allowing them to be interpreted similarly to the p -values used for dichotomous items.

Item difficulty statistics are reviewed as part of the item analysis process to evaluate whether items function as intended and provide an appropriate range of difficulty for the assessment. Items with unusually high or low difficulty may be flagged for further review by test developers and content experts.

Item-Total Correlation

An important indicator of item discrimination is the item-total correlation, which measures the relationship between performance on an individual item and performance on the test as a whole. This statistic is used to evaluate whether test takers who perform well on the overall test also tend to perform well on a particular item.

For *Praxis* tests, item-total correlations are used to assess the extent to which an item functions consistently with the construct measured by the test. Higher positive values indicate that stronger-performing test takers are more likely to answer the item correctly or receive higher scores on the item, whereas lower values suggest that the item may not effectively differentiate among test takers with different levels of proficiency.

Item-total correlations generally range from -1.0 to 1.0 , with positive values indicating that item performance is aligned with overall test performance. Positive item-total correlations are generally expected for operational items. Items with low or negative item-total correlations are reviewed because they may indicate problems such as ambiguous wording, scoring issues, incorrect keys, or content that does not function as intended.

A negative item-total correlation may indicate that higher-performing test takers are less likely to answer the item correctly, or that lower-performing test takers are more likely to answer the item correctly. Such results warrant further review by test developers and content experts to determine whether the item should be revised or removed from scoring.

Omit and Not-Reached Rates

Omit and not-reached rates are used to evaluate item response behavior and to support analyses of test completion and speededness. An item is classified as an *omit* when a test taker views the item, leaves it unanswered, and subsequently answers one or more later items. An item is classified as *not reached* when a test taker leaves the item unanswered and does not respond to any subsequent items. In these cases, the unanswered item and all remaining items at the end of the test are treated as *not reached* for statistical analyses.

Distractor Analyses

Distractor analyses are conducted for selected-response (SR) items to evaluate the effectiveness of incorrect response options. For each item, the proportion of test takers selecting each response option is examined for the full analysis sample and separately for a subgroup of higher-performing test takers. Items may be flagged for review when a substantial number of higher-performing test takers select an incorrect response option rather than the keyed answer. Such patterns may indicate the presence of multiple defensible answers, an incorrect key, or other content-related concerns.

Distractor analyses may also include statistics that evaluate the relationship between selection of a distractor and overall test performance. Distractors that are frequently selected by higher-performing test takers are reviewed because they may indicate ambiguous wording, an incorrect key, or other content issues.

Item Response Theory Analyses

Item response theory (IRT) is a family of statistical models used to describe the relationship between a test taker's level of proficiency and the probability of a particular response to an item. Unlike classical item analyses, IRT models evaluate item performance independently of the specific group of test takers used in the analysis and provide estimates of item characteristics such as difficulty and discrimination.

Within the *Praxis* program, IRT analyses may be used to calibrate items, support score scaling and equating, evaluate item performance, develop and maintain item banks, and support test assembly. These analyses help ensure that scores reported across different forms and administrations have a consistent interpretation.

For SR items, one-parameter or two-parameter logistic IRT models may be used, depending on the characteristics of the test and the intended application. IRT calibrations are conducted using established psychometric software and procedures that support the development, maintenance, and score comparability of *Praxis* tests.

Differential Item Functioning Analyses

Differential item functioning (DIF) analyses are conducted to evaluate whether test items function comparably across groups of test takers with similar levels of overall proficiency. A fundamental assumption of DIF analysis is that individuals who demonstrate comparable performance on the test should have a similar probability of answering an item correctly, regardless of subgroup membership. DIF analyses therefore help identify items for which differences in performance may be attributable to factors other than the knowledge, skills, or abilities the test is intended to measure.

DIF analyses are conducted once sufficient test taker response data have been obtained. Test takers are matched on a criterion measure, typically the test total score excluding the item under review, and item performance is compared across groups. For tests that measure multiple content domains, a more homogeneous content-area score may be used as the matching criterion when appropriate. If testing volumes are insufficient to support DIF analyses after a single administration, data may be accumulated across multiple administrations until adequate sample sizes are available.

DIF results are reviewed as part of the *Praxis* fairness review process. Items showing substantial DIF are evaluated by test developers, psychometricians, and content experts to determine whether the observed differences can be explained by characteristics of the item unrelated to the intended construct. Items found to be influenced by construct-irrelevant factors may be removed from scoring and are generally excluded from future test forms unless required to meet test specifications under exceptional circumstances.

DIF Statistics

DIF analyses for *Praxis* tests are based on the Mantel-Haenszel DIF index expressed on the ETS delta scale (MH D-DIF; Holland & Thayer, 1985)¹. This index measures the extent to which an item is relatively easier or more difficult for one subgroup than another after differences in overall proficiency have been controlled statistically.

When sufficient data are available, DIF analyses are typically conducted for gender and racial or ethnic subgroup comparisons. Examples include male and female test takers, as well as comparisons between White (non-Hispanic) test takers and other racial or ethnic groups. Positive DIF values indicate that the focal group performed better on the item than the matched reference group, whereas negative values indicate that the reference group performed better than the matched focal group. A DIF value of zero indicates that matched test takers in the reference and focal groups performed similarly on the item.

Based on the magnitude and statistical significance of the DIF statistics, items are classified into one of three categories: A (low DIF), B (moderate DIF), or C (high DIF) (Dorans & Holland, 1993). Category C items receive additional review by fairness committees to determine whether the observed DIF reflects construct-irrelevant influences. Items determined to present fairness concerns may be removed from scoring, revised, or excluded from future operational use.

The DIF procedures described above are designed to identify differences in item performance after controlling for overall proficiency.

Test-Form Equating

Praxis tests are administered through multiple test forms, each containing a unique set of SR items, CR items, or both. To ensure that scores from different forms can be interpreted consistently, ETS applies equating procedures in accordance with the *ETS Standards for Quality and Fairness*. Equating statistically adjusts for minor differences in form difficulty so that reported scores have the same meaning regardless of the form administered. Consequently, a given scaled score reflects the same level of performance across all forms of a test. The equating methods used for *Praxis* tests vary according to factors such as test format, testing volume, and score-reporting requirements (von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004). The scaling and equating procedures used for *Praxis* tests are described below.

Scaling

To distinguish reported scores from raw scores, *Praxis* tests use reporting scales. Scaling is the process of converting raw scores to a reporting scale with predetermined lower and upper limits. Most *Praxis* tests use a scaled score range of 100 to 200. The use of a common reporting scale supports consistent interpretation of scores across different forms of the same test.

When a new *Praxis* test is first introduced, an initial raw-score-to-scaled-score conversion is established. Raw scores associated with very low levels of performance are assigned values near the lower end of the reporting scale, while raw scores associated with very high levels of performance are assigned values near the upper end of the scale. Scores falling between these points are converted proportionally to the reporting scale. This initial conversion establishes the reporting scale for the test and provides the foundation for subsequent equating and score reporting.

¹ Delta (Δ) is a transformed measure of item difficulty used in ETS DIF analyses. Lower values correspond to easier items, whereas higher values correspond to more difficult items.

Equating

To maintain score comparability across different forms of the same test, *Praxis* tests use equating procedures to place scores from different forms onto a common reporting scale. Following the initial scaling of a test, each subsequent form is equated to the established score scale prior to score reporting. Equating accounts for differences in form difficulty that may affect score distributions, helping ensure that reported scores have a consistent interpretation regardless of the form administered. Standard test takers with a valid raw score are included in the equating analyses.

The NEAT Design

The Non-Equivalent Groups Anchor Test (NEAT) design is the equating approach most commonly used for *Praxis* tests. Under this design, different groups of test takers take different forms of a test, while a set of common items serves as an anchor linking the forms. The anchor items are selected to be representative of the test's content and statistical characteristics. The performance of test takers on the anchor items provides the basis for adjusting scores so that forms can be placed on a common scale.

The Equivalent Groups Design

For tests with sufficiently large testing volumes, an equivalent groups design may be used. Under this approach, a new form and an existing form are administered during the same testing period, with test takers effectively assigned at random to one of the forms. Because the groups are expected to be comparable in ability, differences in performance can be attributed primarily to differences in form difficulty. Both linear and nonlinear equating methods may be applied under this design.

The Single Group Design

This design is used in situations where a modification to an operational form requires a revised score conversion, such as the removal of an item from scoring. Under this design, the same group of test takers provides data for both score conversions being compared. Equating methods are applied to establish a relationship between the original and revised forms, allowing scores to remain comparable and consistent with previously reported results.

The SiGNET Design

For tests with very low testing volumes, traditional equating designs may not provide sufficient data to support stable equating results. In these cases, the *Praxis* program uses the Single-Group Equating with Nearly Equivalent Test Forms (SiGNET) design. Under this approach, tests are assembled from multiple item clusters, or testlets, that represent the content specifications of the full test. One or more testlets may contain pretest items that are being evaluated for future operational use.

After sufficient response data have been accumulated, a new form is created by replacing an operational testlet with a pretested testlet that has similar content and psychometric characteristics. Single-group equating methods are then used to place the revised form onto the existing score scale. This process allows new content to be introduced gradually while maintaining score comparability across administrations.

Figure 2 illustrates the SiGNET design and the process used to replace operational content with pretested content over time.

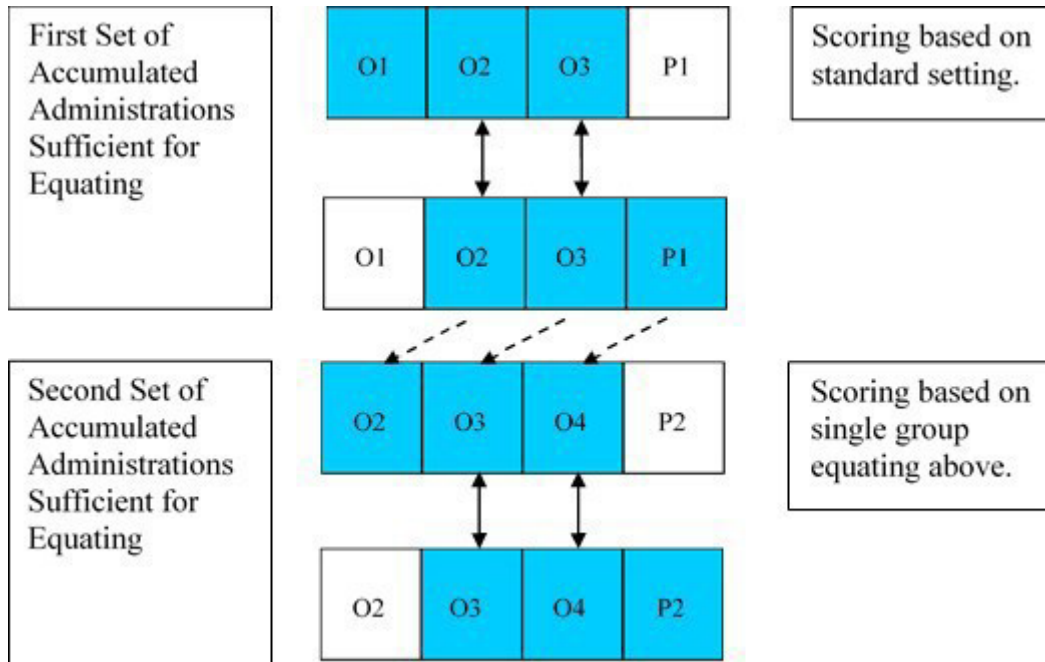


Figure 2. The SiGNET Equating Design

Note: O1–O4 = Operational testlets; P1–P2 = Pretest testlets

The ISD Design

For selected computer-delivered tests with moderate to high testing volumes and no CR items, the Interchangeable Section Design (ISD) with IRT pre-equating may be used. Under this approach, tests are assembled from multiple testlets, each representing a specific content area or a miniature version of the full test. Multiple versions of each testlet are developed with comparable content and statistical characteristics.

During test delivery, operational and pretest testlets are combined to create a form that meets the test specifications. This design supports a large number of possible form combinations, strengthens test security, and provides data for ongoing IRT calibration and pre-equating.

Prior to implementation, sufficient response data are collected to calibrate items and establish the item pool used for testlet assembly. IRT models are then used to estimate item parameters and develop score conversions that support a common reporting scale across different combinations of testlets.

Figure 3 provides an example of the ISD design.

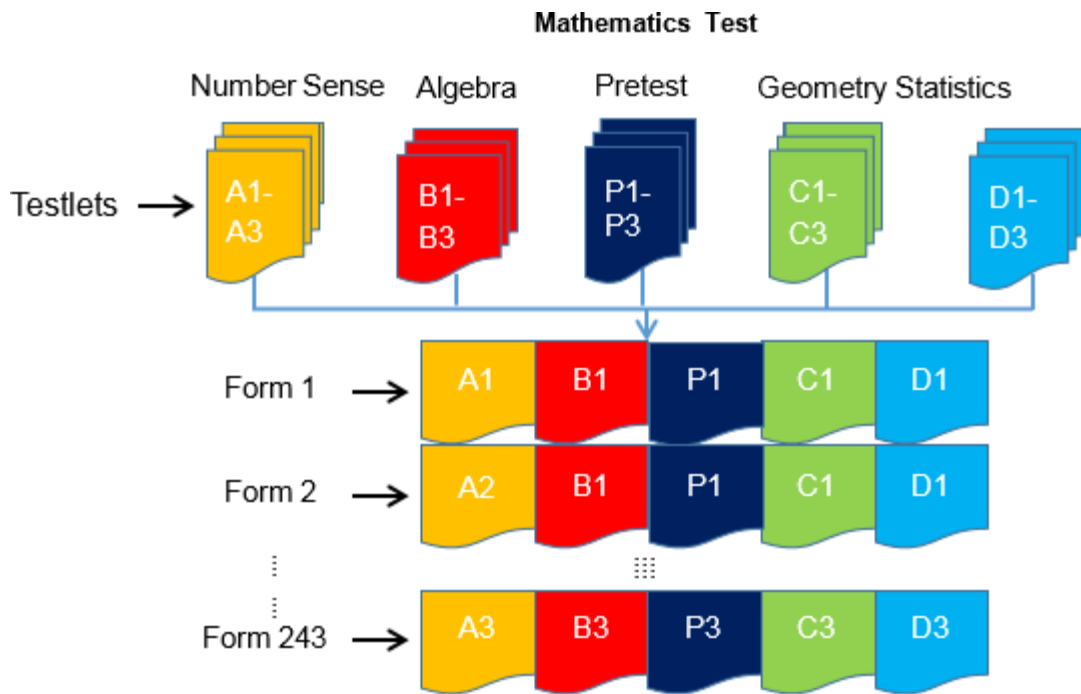


Figure 3. An Example of the ISD Design

IRT-based Linear Form Design

For some linear *Praxis* tests with moderate to high testing volumes, IRT pre-equating may also be applied without using the ISD framework. In these cases, IRT-calibrated item parameters are used to support score conversions and maintain score comparability across test forms. The pre-equating procedures are similar to those described for the ISD design.

Test Scoring

Scaled scores derived through the scaling and equating processes are used for reporting to test takers, institutions of higher education, and other authorized score recipients. These processes help ensure that reported scores have a consistent interpretation and are comparable across different forms and administrations of the same test.

Some *Praxis* tests are available through the *Praxis* Steps program, which provides a modular testing option that allows candidates to demonstrate content mastery across multiple content modules. For *Praxis* Steps tests, each module is independently calibrated and equated to support score comparability. Performance across modules is combined using a weighted average based on the test design to produce a single test-level score for reporting purposes. This approach enables candidates to demonstrate proficiency across the full test while maintaining a common score interpretation regardless of the specific testing pathway used.

Test Statistics

Reliability

Reliability refers to the consistency and stability of test scores. Because any test contains only a sample of the knowledge, skills, or tasks that could be measured, observed scores inevitably contain some degree of measurement error. Reliability provides an indication of the extent to which score differences among test takers reflect true differences in proficiency rather than random sources of variation.

Reliability estimates support the interpretation and use of test scores by indicating how consistently a test measures the intended construct. Higher reliability values indicate that scores are more consistent and less influenced by measurement error.

Reliability estimates are selected according to the psychometric model used for the test. For tests composed entirely of dichotomously scored SR items, reliability is typically estimated using the Kuder-Richardson (1937) Formula 20 (KR-20). For tests that include CR items or other polytomously scored items, reliability is typically estimated using Cronbach's alpha. Both methods provide estimates of internal consistency reliability based on the relationships among item scores.

For tests scored using IRT, marginal reliability is used to estimate the reliability of proficiency estimates. Marginal reliability reflects the precision of the latent trait estimates produced by the IRT model and is based on the variability of proficiency estimates and their associated measurement error (Green et al., 1984).

Standard Error of Measurement

The standard error of measurement (SEM) is an indicator of the amount of measurement error associated with reported scores. It may be interpreted as the expected variation in a test taker's score across repeated administrations of comparable forms of the same test, assuming no meaningful change in proficiency.

The SEM is reported in the same units as the test score and provides information about the precision of score interpretations. Smaller SEM values indicate more precise measurements and greater confidence in reported scores.

Estimates of the SEM are provided for *Praxis* tests in Appendix A. When testing volumes are small, data from multiple administrations may be combined to obtain more stable estimates.

The conditional standard error of measurement (CSEM) extends this concept by providing an estimate of measurement error at specific score levels. Because measurement precision may vary across the score scale, the CSEM can provide more detailed information about the precision of scores for test takers at different performance levels. CSEMs for *Praxis* tests are computed using established psychometric procedures (Lord, 1984).

Reliability of Classification

Because *Praxis* tests are often used to support licensure and certification decisions, the consistency and accuracy of pass/fail classifications are important considerations. The *Praxis* program uses the Livingston and Lewis (1995) methodology to evaluate classification accuracy and classification consistency.

Classification accuracy refers to the extent to which decisions based on observed scores agree with decisions that would be made if a test taker's true proficiency were known. Classification consistency refers to the extent to which the same decision would be reached if the test taker completed an equivalent form of the test.

Classification accuracy and consistency generally increase when the average performance of test takers is well above or well below the passing score. In such situations, fewer test takers score near the decision point, reducing the likelihood that test takers will be classified differently because of measurement error.

Reliability of Scoring

For *Praxis* tests that include CR items, reliability also depends on the consistency of the scoring process. Scoring reliability evaluates the extent to which different qualified scorers assign similar scores to the same responses when applying the established scoring criteria.

Scoring reliability is assessed through analyses of agreement among independent scorers and through estimates of the error associated with the scoring process. These analyses provide an indication of the consistency with which scoring standards are applied across responses and scoring sessions.

Estimates of the standard errors of scoring are reported in Appendix A for tests that include CR items. For tests without CR items, the standard error of scoring is effectively zero because responses are scored electronically using predefined answer keys.

Speededness

Speededness refers to the extent to which test takers are unable to attempt items because insufficient time remains to complete the test. Analyses of speededness help determine whether the time limits are appropriate and whether test scores primarily reflect the intended knowledge and skills rather than the effects of time constraints.

The *Praxis* program evaluates speededness using multiple indicators, including the percentage of test takers who complete all items, the percentage who complete a substantial portion of the test, the number of items reached by most test takers, and the relationship between not-reached responses and overall score variability.

Evidence of speededness may exist when substantial numbers of test takers do not reach the final items of a test. However, the absence of not-reached responses alone does not necessarily indicate that a test is not speeded, because some test takers may have benefited from additional time even if they reached all items. Consequently, speededness is evaluated using multiple sources of evidence, including completion patterns and item-level statistics.

Additional Feedback

For selected *Praxis* tests, test takers may receive information about content areas that may benefit from additional study or development based on their test performance. This feedback is intended to support continued learning and preparation by identifying areas of relative strength and opportunities for additional learning.

For eligible examinees, this feedback may also support participation in the *Praxis* Bridge program. Examinees who do not achieve a passing score but meet specified eligibility criteria may have the

opportunity to demonstrate competency through *Praxis* Bridge learning modules aligned to identified content areas. Test performance information is used to identify the content area most appropriate for targeted learning and support.

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 rubric⁷. The ratings that the scorers assign are based on a rubric developed by educators who are specialists in the subject area.

All scorers receive training before they score operational responses. The score on any single CR test item is the sum of the scores for CR items as assigned by the two scorers.

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. 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 PsychologistsSM(NASPSM)
 - Department of Defense Education Activity (DODEA)
 - Any other entity designated to receive scores by the state or law.

Score Information for States and Institutions

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

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

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

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

Title II Reporting

Overview

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

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

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

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

In this three-stage reporting process:

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

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

Submission of the required institutional and state pass rates is a complex process. For example, while institutions of higher education know the names of program completers, they do not necessarily have complete records of their *Praxis* test scores because students often do not designate their colleges as a score recipient. ETS's Title II services manage the logistical complexities for its clients.

Customized Reporting

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

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

This database system:

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

Client Support

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

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

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

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

Appendix A – 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 29 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 selected-response test is not applicable because selected-response 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	1158	165	154 - 173	163.6	13.3	5.5	n/a	0.83
Algebra I (5162)	100 - 200	1	1364	161	148 - 174	160.2	19.4	7.2	n/a	0.84
American Sign Language (0634)	100 - 200	1	61	160	140 - 160	152.6	19.8	i	i	i
Art: Content Knowledge (5134)	100 - 200	1	5546	163	154 - 171	161.5	14.6	5.6	n/a	0.85
Art: Content and Analysis (5135)	100 - 200	1	2078	165	156 - 173	163.3	13.5	5.8	2.3	0.80
Audiology (5343)	100 - 200	1	2582	169	164 - 176	169.1	10.2	5.7	n/a	0.80
Biology (5236)	100 - 200	1	6933	161	146 - 174	159.2	21.1	5.5	n/a	0.93
Braille Proficiency (0633)	100 - 200	1	18	193.5	184 - 199	189.7	13.3	6.2	c	0.88
Business Education: Content Knowledge (5101)	100 - 200	1	5658	169	159 - 179	168.0	14.8	5.6	n/a	0.89
Chemistry (5246)	100 - 200	1	2271	156	143 - 174	156.5	22.7	6.0	n/a	0.93
Chinese (Mandarin): World Language (5665)	100 - 200	1	346	193	188 - 196	189.0	14.2	4.1	1.5	0.94
Citizenship Education: Content Knowledge (5087)	100 - 200	1	74	159.5	150 - 174	161.2	18.3	5.4	n/a	0.91
Communication and Literacy: Reading (5714)	100 - 200	2	56	168	150 - 182	165.0	23.3	7.5	n/a	0.88
Communication and Literacy: Writing (5724)	100 - 200	2	68	162	148 - 170	158.7	16.8	5.7	1.7	0.81
Computer Science (5652)	100 - 200	1	2146	161	143 - 181	161.1	23.6	6.1	n/a	0.94
Connecticut Administrator Test (6412)	100 - 200	1	1203	170	162 - 176	168.8	10.2	5.7	n/a	0.72
Core Academic Skills for Educators Mathematics (5733)	100 - 200	2	16158	166	150 - 180	163.2	23.0	7.6	n/a	0.90
Core Academic Skills for Educators Reading (5713)	100 - 200	2	15452	170	158 - 184	169.2	18.7	7.5	n/a	0.88
Core Academic Skills for Educators Writing (5723)	100 - 200	2	16696	162	154 - 168	160.3	13.4	5.7	1.7	0.81
Early Childhood Education (5025)	100 - 200	1	12824	167	156 - 178	164.7	17.7	5.5	n/a	0.90
Early Childhood Education Foundational Knowledge (5533)	100 - 200	1	149	160	149 - 171	159.2	14.9	5.8	n/a	0.81
Early Childhood Education Foundational Knowledge and Content (5534)	100 - 200	1	559	166	157 - 173	164.3	13.7	5.8	n/a	0.82
Early Childhood: Math and Science (5028)	100 - 200	1	532	171	161 - 183	169.2	18.5	6.5	n/a	0.87
Early Childhood: Reading and Language Arts and Social Studies (5027)	100 - 200	1	527	174	163 - 184	171.8	16.2	6.6	n/a	0.81

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Earth and Space Sciences (5572)	100 - 200	1	1382	164	151 - 179	162.8	20.5	5.4	n/a	0.93
Economics (5911)	100 - 200	1	353	152	141 - 167	153.5	20.9	6.2	n/a	0.91
Education of Young Children (5024)	100 - 200	1	4147	166	158 - 175	164.5	14.8	5.4	1.9	0.83
Educational Leadership: Administration and Supervision (5412)	100 - 200	1	14058	167	159 - 174	165.8	11.1	5.7	n/a	0.78
Elementary Education Fundamentals: Mathematics (8003)	100 - 200	1	147	159	145 - 174	158.8	20.1	i	n/a	i
Elementary Education Fundamentals: Reading and Language Arts (8002)	100 - 200	1	154	163.5	152 - 173	162.3	13.4	i	n/a	i
Elementary Education Fundamentals: Science (8005)	100 - 200	1	134	155	137 - 173	154.1	23.0	i	n/a	i
Elementary Education Fundamentals: Social Studies (8004)	100 - 200	1	141	151	135 - 165	151.0	18.1	i	n/a	i
Elementary Education Fundamentals: Teaching Reading (8006)	100 - 200	1	f	f	f	f	f	f	n/a	f
Elementary Education: Content Knowledge (5018)	100 - 200	1	1841	164	143 - 176	158.2	22.8	5.7	n/a	0.91
Elementary Education: Curriculum, Instruction, and Assessment (5017)	100 - 200	1	4459	163	154 - 173	161.8	16.1	6.1	n/a	0.87
Elementary Education: Math Specialist (5037)	100 - 200	1	127	175	162 - 184	171.1	18.5	6.9	n/a	0.85
Elementary Education: Math and Science (5008)	100 - 200	1	5121	165	158 - 176	164.4	17.9	6.9	n/a	0.85
Elementary Education: Mathematics (7003)	100 - 200	1	6755	165	148 - 182	162.9	23.2	9.2	n/a	0.84
Elementary Education: Mathematics Subtest (5003)	100 - 200	1	37961	169	155 - 183	166.4	22.5	9.2	n/a	0.85
Elementary Education: Mathematics-CKT (7813)	100 - 200	1	12289	158	146 - 170	156.0	19.2	7.6	n/a	0.82
Elementary Education: Reading and Language Arts Subtest (5002)	100 - 200	1	36888	168	159 - 178	166.7	16.1	7.1	n/a	0.83
Elementary Education: Reading and Language Arts-CKT (7812)	100 - 200	1	4242	168	161 - 177	167.2	14.4	6.8	n/a	0.76

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: Reading, Language Arts and Social Studies (5007)	100 - 200	1	4872	170	162 - 179	170.1	13.8	6.2	n/a	0.80
Elementary Education: Science (7005)	100 - 200	1	6537	164	151 - 176	163.0	18.9	8.2	n/a	0.80
Elementary Education: Science Subtest (5005)	100 - 200	1	37482	167	158 - 179	166.2	17.8	8.2	n/a	0.80
Elementary Education: Science-CKT (7814)	100 - 200	1	4107	171	159 - 185	170.1	18.5	8.0	n/a	0.80
Elementary Education: Social Studies (7004)	100 - 200	1	6671	158	147 - 172	159.7	18.9	8.0	n/a	0.81
Elementary Education: Social Studies Subtest (5004)	100 - 200	1	37857	164	154 - 177	164.4	18.2	8.0	n/a	0.81
Elementary Education: Social Studies-CKT (7815)	100 - 200	1	4376	169	162 - 179	168.1	15.7	7.3	n/a	0.74
Elementary Education: Teaching Reading (7002)	100 - 200	1	8840	155	145 - 164	154.0	13.9	6.5	n/a	0.80
Elementary Education: Three Subject Bundle-Mathematics (5903)	100 - 200	1	7760	169	155 - 184	167.4	22.2	9.2	n/a	0.85
Elementary Education: Three Subject Bundle-Science (5905)	100 - 200	1	6141	169	158 - 179	166.9	18.4	8.2	n/a	0.80
Elementary Education: Three Subject Bundle-Social Studies (5904)	100 - 200	1	6204	165	153 - 179	164.9	19.1	8.0	n/a	0.81
English Language Arts: Content Knowledge (5038)	100 - 200	1	16180	176	168 - 184	174.5	13.9	5.0	n/a	0.90
English Language Arts: Content and Analysis (5039)	100 - 200	1	4343	173	165 - 179	170.6	13.3	4.7	2.2	0.82
English to Speakers of Other Languages (5362)	100 - 200	1	16098	175	165 - 183	173.7	13.5	5.2	n/a	0.86
Family and Consumer Sciences (5123)	100 - 200	1	1035	168	160 - 173	166.0	11.2	5.1	n/a	0.75
French: World Language (5174)	100 - 200	1	646	170	158 - 181	169.0	17.0	4.9	2.0	0.93
Fundamental Subjects: Content Knowledge (5511)	100 - 200	1	7306	169	156 - 183	168.5	18.7	5.9	n/a	0.92
General Science (5436)	100 - 200	1	5034	155	139 - 173	154.6	24.8	6.2	n/a	0.94
Geography (5921)	100 - 200	1	225	171	158 - 183	169.0	18.7	5.1	n/a	0.90
Geometry (5163)	100 - 200	1	229	148	129 - 161	146.3	21.2	8.7	n/a	0.86
German: World Language (5183)	100 - 200	1	149	180	157 - 200	174.6	24.9	5.2	2.1	0.96

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Gifted Education (5358)	100 - 200	1	2538	163	158 - 169	162.2	10.0	5.0	n/a	0.74
Government/Political Science (5931)	100 - 200	1	611	166	157 - 176	165.2	16.1	5.2	n/a	0.92
Health Education (5551)	100 - 200	1	2962	164	155 - 173	162.8	13.8	5.1	n/a	0.86
Health Occupations (5052)	100 - 200	1	53	177	169 - 187	178.3	11.3	i	n/a	i
Health and Physical Education: Content Knowledge (5857)	100 - 200	1	7924	163	154 - 171	162.1	14.1	5.6	n/a	0.82
Interdisciplinary Early Childhood Education (5023)	100 - 200	1	841	176	168 - 184	175.0	11.9	4.6	n/a	0.75
Japanese: World Language (5661)	100 - 200	1	40	198.5	164 - 200	181.8	24.8	i	i	i
Journalism (5224)	100 - 200	1	174	181	171 - 187	178.4	11.9	4.9	n/a	0.86
Latin (5601)	100 - 200	1	61	185	162 - 200	179.2	20.7	4.9	n/a	0.95
Marketing Education (5561)	100 - 200	1	628	164	154 - 175	162.9	17.0	5.7	n/a	0.90
Mathematics (5165)	100 - 200	1	14251	169	151 - 186	165.6	25.1	7.0	n/a	0.92
Middle School English Language Arts (5047)	100 - 200	1	6796	162	150 - 169	159.2	14.5	5.6	2.3	0.79
Middle School Mathematics (5164)	100 - 200	1	12711	165	151 - 180	163.3	21.2	7.1	n/a	0.88
Middle School Science (5442)	100 - 200	1	7413	160	146 - 175	158.9	21.1	5.9	n/a	0.93
Middle School Social Studies (5089)	100 - 200	1	4742	164	150 - 178	163.2	20.3	6.2	2.3	0.88
Middle School Social Studies (5589)	100 - 200	1	406	157	143 - 172	156.1	21.0	5.4	n/a	0.93
Music: Instrumental and General Knowledge (5115)	100 - 200	1	650	161	152 - 172	160.4	16.0	5.9	n/a	0.89
Music: Vocal and General Knowledge (5116)	100 - 200	1	448	163	153 - 173	161.1	17.3	5.8	n/a	0.88
Music: Content Knowledge (5113)	100 - 200	1	4586	163	155 - 173	162.9	14.6	5.7	n/a	0.85
Music: Content and Instruction (5114)	100 - 200	1	1999	163	151 - 170	160.1	14.4	6.0	1.7	0.77
ParaPathways: Mathematics (5759)	310 - 350	1	2425	340	332 - 345	337.8	9.4	3.8	n/a	0.85
ParaPathways: Reading and Writing (5758)	310 - 350	1	2489	332	325 - 338	331.2	8.5	3.5	n/a	0.83
ParaPro Assessment (1755)	420 - 480	1	92450	468	459 - 474	465.5	11.6	3.4	n/a	0.94

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Pennsylvania Grades 4-8 Core Assessment: English Language Arts and Social Studies (5154)	100 - 200	1	3392	159	149 - 171	159.1	18.6	8.2	n/a	0.81
Pennsylvania Grades 4-8 Core Assessment: Mathematics and Science (5155)	100 - 200	1	3379	169	156 - 182	167.3	21.3	8.4	n/a	0.83
Pennsylvania Grades 4-8 Core Assessment: Pedagogy (5153)	100 - 200	1	1767	177	169 - 184	176.4	11.2	5.7	n/a	0.78
Pennsylvania Grades 4-8 Subject Concentration: English Language Arts (5156)	100 - 200	1	811	165	154 - 178	163.8	19.3	6.9	n/a	0.87
Pennsylvania Grades 4-8 Subject Concentration: Mathematics (5158)	100 - 200	1	1112	174	156 - 183	168.7	21.1	7.7	n/a	0.85
Pennsylvania Grades 4-8 Subject Concentration: Science (5159)	100 - 200	1	568	162	148 - 174	160.0	19.2	6.7	n/a	0.86
Pennsylvania Grades 4-8 Subject Concentration: Social Studies (5157)	100 - 200	1	485	163	150 - 175	161.8	20.3	7.1	n/a	0.88
Physical Education: Content Knowledge (5091)	100 - 200	1	7075	153	148 - 159	152.9	9.3	4.0	n/a	0.85
Physical Education: Content and Design (5095)	100 - 200	1	2135	169	161 - 174	166.0	12.6	5.5	2.3	0.78
Physical Science (5485)	100 - 200	1	261	149	131 - 163	147.7	21.4	6.1	n/a	0.91
Physics (5266)	100 - 200	1	1341	164	147 - 182	162.4	23.9	5.6	n/a	0.95
Principles of Learning and Teaching: 5-9 (5623)	100 - 200	1	4027	174	166 - 182	173.3	11.4	5.0	2.1	0.82
Principles of Learning and Teaching: 7-12 (5624)	100 - 200	1	17077	176	167 - 183	174.2	12.5	5.4	2.3	0.83
Principles of Learning and Teaching: Early Childhood (5621)	100 - 200	1	4345	167	159 - 175	165.8	12.2	5.0	2.1	0.80
Principles of Learning and Teaching: K-6 (5622)	100 - 200	1	19055	174	166 - 182	173.2	11.5	4.9	2.0	0.85
Principles of Learning and Teaching: Pre K-12 (5625)	100 - 200	1	5251	166	158 - 176	165.8	14.5	6.5	n/a	0.83
Professional Teaching Knowledge Exam (5626)	100 - 200	1	i	i	i	i	i	i	n/a	i
Psychology (5391)	100 - 200	1	323	167	156 - 178	166.6	16.2	5.1	n/a	0.90
Reading Specialist (5302)	100 - 200	1	3995	172	165 - 178	171.0	10.1	4.8	1.9	0.81
Russian: World Language (5671)	100 - 200	1	11	165	137 - 178	157.1	23.2	i	i	i

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
STEM for the Elementary Grades (5036)	100 - 200	1	564	185	180 - 192	184.8	10.0	5.6	n/a	0.55
School Counselor (5422)	100 - 200	1	10647	165	158 - 172	163.8	11.4	5.1	n/a	0.85
School Leaders Licensure Assessment (6990)	100 - 200	1	16294	170	162 - 177	169.0	11.5	5.1	1.8	0.82
School Librarian (5312)	100 - 200	1	3416	171	163 - 179	169.8	12.7	5.5	n/a	0.84
School Psychologist (5403)	100 - 200	1	8004	172	165 - 178	171.3	9.9	4.8	n/a	0.79
School Superintendent Assessment (6991)	100 - 200	1	2455	173	166 - 180	172.4	11.1	5.2	1.0	0.84
Social Studies (5581)	100 - 200	1	6119	165	153 - 177	164.2	18.2	5.0	n/a	0.92
Social Studies: Content and Interpretation (5086)	100 - 200	1	2233	158	147 - 169	156.8	18.4	5.7	2.1	0.87
Sociology (5952)	100 - 200	1	136	171	160 - 180	169.2	14.7	4.9	n/a	0.87
Spanish: World Language (5195)	100 - 200	1	4543	177	164 - 187	172.9	18.6	5.2	2.3	0.93
Special Education: Early Childhood/Early Intervention (5692)	100 - 200	1	2347	170	161 - 177	168.7	12.9	5.8	n/a	0.81
Special Education: Education of Deaf and Hard of Hearing Students (5272)	100 - 200	1	426	166	157 - 172	163.0	13.6	5.5	n/a	0.81
Special Education: Foundational Knowledge (5355)	100 - 200	1	27003	171	161 - 179	169.3	13.1	5.3	n/a	0.85
Special Education: Severe to Profound (5547)	100 - 200	1	2294	175	166 - 182	173.4	12.3	5.1	n/a	0.82
Special Education: Teaching Speech to Students with Language Impairments (5881)	100 - 200	1	54	152	149 - 161	153.5	12.7	5.8	n/a	0.84
Special Education: Teaching Students with Behavioral Disorders/Emotional Disturbances (5372)	100 - 200	1	186	172	161 - 182	170.2	15.6	4.0	n/a	0.90
Special Education: Teaching Students with Intellectual Disabilities (5322)	100 - 200	1	78	177	166 - 185	173.6	15.5	4.9	n/a	0.86
Special Education: Teaching Students with Learning Disabilities (5383)	100 - 200	1	347	165	154 - 173	163.4	15.2	5.2	n/a	0.88
Special Education: Teaching Students with Visual Impairments (5282)	100 - 200	1	376	169	164 - 175	168.2	11.1	5.4	n/a	0.77
Speech Communication: Content Knowledge (5221)	100 - 200	1	488	158	147 - 169	156.9	15.4	4.7	n/a	0.87
Speech and Theatre (5222)	100 - 200	1	105	158	148 - 166	156.0	14.5	5.9	n/a	0.84
Speech-Language Pathology (5331)	100 - 200	1	32094	174	167 - 181	173.8	10.5	5.1	n/a	0.85

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Teaching Reading K-12 (5206)	100 - 200	1	899	165	157 - 172	163.3	12.0	4.4	1.4	0.86
Teaching Reading: Elementary (5205)	100 - 200	1	40294	168	161 - 176	167.3	12.9	4.7	1.6	0.89
Technology and Engineering Education (5053)	100 - 200	1	655	169	160 - 177	168.1	13.0	5.3	n/a	0.82
Theatre (5641)	100 - 200	1	826	168	160 - 176	167.1	13.3	5.4	n/a	0.84
World Languages Pedagogy (5841)	100 - 200	1	486	172	161 - 183	171.0	16.4	7.3	2.2	0.83
World and U.S. History: Content Knowledge (5941)	100 - 200	1	3010	159	147 - 171	158.1	18.7	5.5	n/a	0.91

Notes:

“Number of Test Takers,” “Median,” and “Average Performance Range” were calculated from the records of test takers who took the test between August 2023 and July 2026, 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, juniors, and earned associates degree.

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.

n/a = not applicable

Bibliography

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.
- Dorans, N., & Holland, P. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. Holland and H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355–368.
- Educational Testing Service, *ETS Guidelines for Fairness Review of Assessments*, Princeton, N.J., 2009.
- Educational Testing Service, *ETS Standards for Quality and Fairness*, Princeton, N.J., 2014
- Educational Testing Service, *Questions to Ask About Teacher Testing*, Princeton, N.J., 2004
- Educational Testing Service, *Proper Use of the PRAXIS Assessments and Related Assessments*, Princeton, N.J., 2016
- Holland, P.W. & Thayer, D.T. (1985). An alternative definition of the ETS delta scale of item difficulty (RR-85-43). Princeton, N.J.: Educational Testing Service.
- Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), *Test Validity*, pp. 129–145. Hillsdale, N J: Lawrence Erlbaum Associates.
- Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Joint Committee on Testing Practices (2005). *Code of Fair Testing Practices in Education*. Washington, D.C.
- Knapp, J., & Knapp, L. (1995). Practice analysis: Building the foundation for validity. In J.C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 93–116). Lincoln, NE: Buros Institute of Mental Measurements.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices* (2nd Ed.). New York: Springer-Verlag.
- Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Livingston, S.A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, 21, 239–243.
- Raymond, M.R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education*, 14, 369–415.

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

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

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

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

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