The PRAXIS Series™

Listening. Learning. Leading.®

Praxis™ Technical Manual

Revised December 2010

Table of Contents

Preface

Purpose of This Manual

The purpose of the *Praxis*™ Technical Manual is to explain:

- The purpose of the *Praxis* tests
- How states use the *Praxis* tests
- The approach ETS® takes in developing *Praxis* tests
- The validity evidence supporting *Praxis* test score use
- How states adopt *Praxis* 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 *Praxis* tests are developed and scored
- Interested in the statistical characteristics of *Praxis* tests

Purpose of *The Praxis Series*™ 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 Series*™ assessments. The *Praxis* tests provide states with testing tools and ancillary services that support their teacher licensing process[1]. These tools include tests of basic academic competency and subject-specific assessments related to teaching.

All states have an abiding interest in ensuring that teachers have the necessary knowledge and skills before they receive a license. To aid states in this effort, the *Praxis* tests assess a test taker's knowledge of important content and skills required to be licensed to teach. States adopt the *Praxis* tests as one measure of helping to ensure that teachers have achieved a specified level of mastery of academic skills, subject area knowledge, and pedagogical knowledge before they grant a teaching license.

Each of *The Praxis Series* tests reflects what practitioners in that field across the United States believe to be important for new teachers. The knowledge and skills measured by the tests are informed by this national perspective, as well as by the content standards recognized by that field. The nexus of these sources of knowledge and skills means that the *Praxis* assessments offer states a unique opportunity to understand if their candidates are meeting the expectations of the profession. The *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 means the same as a person who takes the same *Praxis* test in another state.

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

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

[1] Some states use the term "certification" instead of "licensing."

The *Praxis I*® Tests — Basic Academic Competency

The *Praxis I* tests are designed to measure basic competency in reading, writing, and mathematics. The tests can be taken on paper or computer. Many colleges, universities, and other institutions use the results of *Praxis I* tests primarily as a way of evaluating test takers for entrance into teacher education programs. Many states use the tests in conjunction with *Praxis II*® tests as part of the teacher licensing process.

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

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

States use the *Praxis II* tests for initial teacher licensing as well as throughout the different stages of a teacher's career.

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

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

When a state is considering the adoption of one or more *Praxis II* tests, state representatives are closely involved in many phases of development. For example, teachers of that content area and members of the state department of education are involved in evaluating the relevancy of the test content for entry-level teachers in that state. If a test is adopted, teachers and members of the state department are involved in the process of determining the appropriate passing score for the test. Teachers also are involved in development activities, such as writing and reviewing test items and serving on test committees. Input from teachers and departments of education, therefore, is vital in the development of the *Praxis* tests.

How *The Praxis Series* Assessments Address States' Needs

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

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

- How the *Praxis* tests align with state and national content standards.
- How the *Praxis* tests measure a full range of teaching skills and content knowledge identified in the No Child Left Behind Act.
- 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 Series* program has available many helpful test preparation tools**.** These materials take many forms:

- Study guides and practice tests, some in downloadable eBook format
- Test specifications and sample items available online, free to all candidates
- Faculty and train-the-trainer workshops for school districts and teacher educators to assist them in helping candidates prepare for the test.

Finally, states have a strong interest desire to support the state institutions of higher education that prepare teachers. The *Praxis* program provides workshops for higher education faculty to learn about *Praxis* tests, including how to help their students master the test material. Each year, institutions also receive annual summary reports of their *Praxis* test takers' scores. Finally, the *Praxis* tests offer an additional Title II Reporting Service to institutions of higher education to help them satisfy federal reporting requirements.

Assessment Development

Fairness in Test Development

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

For more explicit guidelines used in item development and review, please see the *ETS Fairness Review Guidelines*.

Test Development Standards

During the *Praxis* test development process, the program follows the strict guidelines detailed in *Standards for Educational and Psychological Testing*:

- 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 defined by the job analysis
- Develop specifications for item types and numbers of items needed to adequately sample the domains of knowledge validated by the job analysis survey
- Develop test items that provide evidence of the measurable-behavior indicators detailed in the test specifications
- Review test items and assembled test forms so that each item has a single best defensible answer and assesses content that is job relevant
- Review test items and assembled forms for potential fairness or bias concerns, overlap, and cueing, revising or replacing items as needed to meet standards. (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.)

How New Tests Are Chosen

Overview

The *Praxis* program provides tests to more than 40 states, U.S. jurisdictions, and professional associations as part of these groups' teacher, professional, and administrator certification processes.

ETS often receives requests to revise an existing test or create a new test to meet state-specific needs. To help evaluate these requests, ETS and the client consider various aspects of the request, including the following areas:

1. The entity (state, agency, etc.) making the request
2. The state's political climate and potential political factors that are influencing the request
3. The certification or licensure area that is being revised. For example, does the new test include a grade-level change?
4. Whether other states also might want to adopt the test
5. Whether the test will be used only for licensure, only for highly qualified status, or for both
6. The projected number of persons expected to be licensed in this area annually
7. The timeline for introducing the test and having a passing score in place
8. Whether the state uses current *Praxis* test(s) and the state's satisfaction level with those tests.
9. Whether the test needs to be aligned with state standards, cost structure, new licensure area, or test administration schedule
10. The test format (MC, CR, MC/CR), test length, and delivery mode

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*, AERA, APA, NCME, 1999).

A test developed to inform licensure[2] decisions is intended to convey the extent to which the test taker (candidate for the credential) has a sufficient level of knowledge and/or skills to perform important occupational activities in a safe and effective manner *(Standards for Educational and Psychological Testing*, 1999). "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, p. 4). 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

[2] Licensure and certification tests are referred to as credentialing tests by the *Standards for Educational and Psychological Testing* (1999). Unless quoted from the Standards, we use the term "licensure."

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*, 1999). 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*, 1999, p. 161). 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*, 1999).

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. Job analysis is the process used to define occupational knowledge and/or skills.

Within the test development cycle, the items in the *Praxis I* and *Praxis II* assessments are developed using an evidence-centered design process (ECD) that adds to the validity 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 through a job analysis (see "Job Analysis" on page 15). Building on this information, test developers then work with the National Advisory Committee, 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.

Job Analysis

The *Standards for Educational and Psychological Testing* (1999) makes it clear that a job analysis needs to be performed to support the content evidence of the validity of a licensure test: "Some form of job or practice analysis provides the primary basis for defining the content domain [of the credentialing test]." A job analysis (known also as practice analysis or role delineation study) refers to a variety of systematic procedures designed to provide a description of occupational

[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.

tasks/responsibilities and/or the knowledge, skills, and abilities believed necessary to perform those tasks/responsibilities. *Praxis I* and *Praxis II* tests use a job analysis process as follows:

- A review of available professional literature and disciplinary (content) standards to develop a draft domain of knowledge and/or skills
- Meetings with a National Advisory Committee of experts to review and revise the draft domain
- A survey of the profession to confirm the importance of the committee-revised domain (see, for example, Knapp and Knapp, 1995; Raymond, 2001; Tannenbaum and Rosenfeld, 1994).

Job analyses are periodically reviewed and revised. A list of ETS job analyses can be found in "Appendix A – *Praxis* Job Analyses."

In ETS job analyses, the committee that is formed:

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

Such diversity and representation reinforces the development of domain knowledge and/or skills that is applicable across the profession. The involvement of various subgroups of experts also is part of the process of developing a test that is considered fair and reasonable to subgroups of practitioners and test takers. The committee's charge is basically to review and revise the draft domain so that it adequately defines the knowledge and/or skills important for safe and effective entry-level practice.

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

Basic analyses (means and standard deviations) are conducted to summarize and interpret responses. Analyses are conducted for the total group of respondents and for subgroups of respondents. The purpose of these analyses is to identify those knowledge and/or skill statements that have been judged to be most important for entering teachers; the subgroup analysis is used to identify content that may not be similarly valued by different groups of educators. The results of the survey are used to inform the development of test content specifications that serve as the blueprint for formal test development. It is this alignment between job analysis outcomes and test content specifications and, ultimately, between test content specifications and test items, that serves to reinforce the valid use of test scores for licensure purposes.

Validity Maintenance

The content covered by a *Praxis I* or *Praxis II* licensure test is evaluated on a periodic basis so that it accurately reflects the current state of knowledge and/or skill requirements of that profession. One rule of thumb is to review the test content every five years; however, some areas—such as those that

are technology-based—may need to be reviewed on a more frequent cycle to keep pace with the changes in the profession.

The critical threshold, however, is crossed if any changes are believed to affect what entering teachers need to know or be able to do for safe and effective practice; that is, any changes that directly relate to the purpose of licensure. Changes in knowledge and/or skills that do not impact expectations of safe and effective practice need not be acted upon unless, of course, the exclusion of such knowledge and/or skill compromises the acceptance of the test by the profession. That is, the test content appears dated and, therefore, no longer seems credible to the profession. The current schedule calls for approximately 20 percent of the *Praxis* tests to be re-evaluated each year so that all tests are examined at least once in a five-year period.

Content reviews are conducted by National Advisory Committees. A review of the test is conducted each time a committee meets. The committee members consider the test content in light of their understanding of the current state of the profession, changes in disciplinary (content) standards, and their experiences. If, in the judgment of the committee and/or test development specialists, relatively substantial modifications to the test content are needed—changes that call into question the alignment of the content domain measured by the existing content and the proposed domain—a survey of the profession to verify the proposed changes is conducted. The results of such a survey could result in the design of a completely new test.

Test Development Process

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

- Research national, state, and professional standards and curricula to verify alignment with the claims made for the test and the test takers.
- Recruit and convene a National Advisory Committee (NAC) to help develop the job analysis claims.
- Conduct job analysis/content validation survey.
- Reconvene the NAC to develop test specifications and blueprints, using the results of the job analysis survey.
- Recruit expert practitioners, who teach the potential test takers and understand the job defined in the job analysis, to write items for the test.
- Develop sufficient numbers of test items to form a pool from which parallel forms can be assembled.
- Review the items developed by trained writers, applying and documenting *ETS Standards for Fairness and Quality* and editorial guidelines. Item reviews also are done by practitioners in the field who may not be trained writers but who have the content expertise to judge the accuracy of the items.
- Prepare the approved test items for publication and assemble them into operational forms.
- 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.
- Print test books and perform final quality-control checks, according to the program's standard operating procedures.
- Administer a pilot test if it is included in the development plan.

- Analyze and review test data from the pilot or first administration to verify that items are functioning as intended and present no concerns about the intended answers or impact on subgroups.

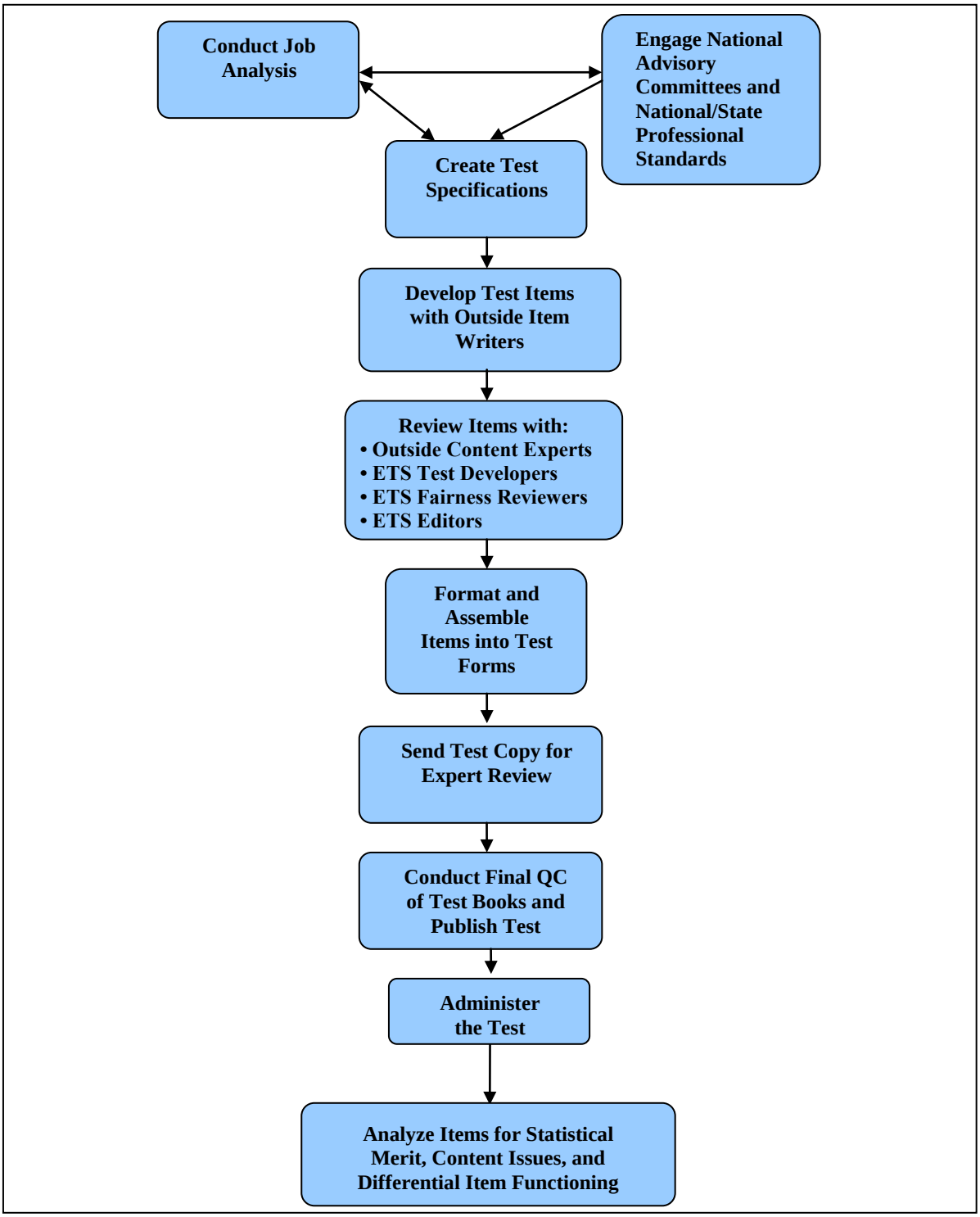

Figure 1: Test Development Process

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

Conduct Job Analysis.

Job analysis is a systematic process, the goal of which is to determine the knowledge and/or skills important for safe and effective entry-level practice. As explained in "Validity," a job analysis is the primary source of validity evidence for licensure tests (*Standards for Educational and Psychological Testing*, 1999). One phase of the job analysis is the construction of a domain of job-related knowledge and/or skills. This is done with the assistance of a National Advisory Committee. The NAC-approved domain is administered as a survey to a large sample of teachers and college faculty for verification of the judged importance of the knowledge and/or skills for entry-level practice. The outcomes of the survey are then used by the NAC to develop test content specifications.

Engage National Advisory Committees and National/State Professional Standards

The National Advisory Committee (NAC) is a group of approximately 15 experts in the field—practicing teachers, teacher educators, and administrators—nominated by state departments of education, professional organizations, deans, superintendents, and colleagues to participate in developing job-related test content specifications. The specifications are necessary to support the validity of licensure test score use (*The Standards for Educational and Psychological Testing*, 1999*)*.

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

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

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

During the second NAC meeting, the results of the survey are presented and discussed. The focus of the presentation is on those knowledge and/or skill statements that were verified by the survey respondents as being important for entry-level practice. The NAC, under the guidance of ETS test developers, uses this information to construct the test content specifications.

Create Test Specifications

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

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

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

Develop Test Items with Outside Item Writers

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

Review Items (by Outside Content Experts, ETS Test Developers, ETS Fairness Reviewers, ETS Editors)

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

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

Format and Assemble Items into Test Forms

When items are ready to be used in a test form, they are formatted for use by a team of experts who are specially trained in layout and formatting procedures. Formatted items selected to be used in a form are assembled in a spreadsheet, using test-assembly software, and automatically laid out in a mock-up of a test book. The auto test layouts are modified as needed and checked for quality by a test layout specialist. These layouts are then checked for quality by test developers, content experts, and test coordinators to verify that the standards documented in the program's Standard Operating Procedures (SOPs) have been met. Changes to each version of the layout are documented in an electronic assembly unit record.

Send Test Copy for Expert Review

Before a test is certified by test developers and the test coordinator as ready to be printed, it receives a content review to verify that every item has a single best answer, which can be defended, and that no item has more than one possible key. The reviewer must understand the purpose of the test and be prepared to

challenge the use of any item that is not important to the job of the beginning practitioner or is not a match to the test specifications.

Conduct Final QC of Test Books and Publish Test

When the reviews of a particular test form have been examined, test developers perform multiple checks of the reviewers' keys against the official key. They must revise the layout, if necessary, and document the revisions. They must certify the test as ready for packaging; that is, it is okay to print. They do this by initialing each page of a copy of the most recent version, and by signing and dating certain pages of the copy. The test coordinator then checks that all steps specified in the SOPs have been followed and signs off on the test copy that will be printed. The certified test book copy is sent to test publishing and the electronic file is archived.

Administer the Test

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

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

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

If it is decided not score an item, a Problem Item Notice (PIN) is issued and distributed. The distribution of a PIN triggers actions in the Statistical Analysis, Assessment Development, and Score Key Management organizations. As a result, official test keys must be updated, items in databases must be revised or deactivated, open reports on flagged items must be reviewed and closed, and the number of items used to compute and report scores must be adjusted.

If there is sufficient test taker volume, Differential Item Functioning (DIF) analyses are run on a new test form to determine if subgroup differences in performance may be due to factors other than the abilities the test is intended to measure. These procedures are described more fully in "Differential Item Functioning (DIF) Analyses" on page 33, and in Holland and Wainer (1993). A DIF panel of content experts decides if items with statistically high levels of DIF (C-DIF) should be dropped from scoring. If that is the case, test developers must prepare a do-not-score PIN and close a report using test creation software. Test developers are responsible for ensuring that C-DIF items are not used in future editions of the test.

Review Processes

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

ETS Standards for Quality and Fairness

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

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

ETS Fairness Review

The *ETS Fairness Review Guidelines* identify 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 I* Tests

Teacher Licensure. The *Praxis I* assessments may be used by the licensing body or agency within a state for teacher licensing decisions. The *Praxis* program requires that before adopting a test, the licensing body or agency must review 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." More than one approach to setting a passing score may be used. The choice of approach should be consistent with the design and format of the test, as should the decision that is made on the basis of the test score.

Entrance into Teacher Preparation Programs. These tests also may be used by institutions of higher education to identify students ("rising juniors") with sufficient reading, writing, and mathematics skills to enter a teacher preparation program. If an institution is in a state that has authorized the use of the *Praxis I* 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 I* scores, when making final entrance decisions.

If an institution of higher education is in a state that has not authorized use of the *Praxis I* 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 II* Tests

Teacher Licensure and NCLB Highly Qualified Compliance. The *Praxis II* tests 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 *Praxis II* tests also may be used by states to satisfy federal No Child Left Behind (NCLB) regulations regarding Highly Qualified teachers. However, before a *Praxis II* test may be used for teacher licensure or for NCLB Highly Qualified compliance, the licensing body or agency must verify that the content of the test is appropriate (valid) for such uses in that particular state. Upon such verification, the licensing body or agency must then establish a passing standard, or cut score. ETS' interpretation of the NCLB regulations is that cut scores for licensure and Highly Qualified status are interchangeable. If a cut score already exists for either purpose, it can be applied to the other.

Program Quality Evaluation. Institutions of higher education may want to use *Praxis II* test 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 II* content test scores as one criterion for permitting students to move on to the clinical portion of their program: the student teaching phase. The use of the *Praxis II* test 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 Series* tests that only focus on pedagogical knowledge (i.e., the Principles of Learning and Teaching set of assessments) should not be used as prerequisites for student teaching.

The *Praxis* program suggests that institutions analyze the content knowledge a candidate must have to perform satisfactorily in the role of student teacher. This is analogous to conducting a small-scale job or practice analysis. The program can then review the *Praxis II* test specifications to verify that it adequately covers the content that is important for its student teachers. If the institution's state does not require that students pass this content test for state licensure, the institution will need to conduct a standard-setting study to establish a minimum score for entrance into student teaching.

There are three scenarios involving the use of *Praxis II* 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 II* content test for licensure, but not as a prerequisite for student teaching; and (3) The state requires the identified *Praxis II* content test neither for licensure nor as a prerequisite for student teaching.

If an institution is in a state that uses the identified *Praxis II* content assessment for licensure, the state may also require candidates 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 *Praxis II* test 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 II* content test for teacher licensure, the institution should review the test specifications to confirm that

the content covered is a necessary prerequisite for entrance into student teaching and that the curriculum which students were exposed to covered that content.

The following summarizes this process:

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

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

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

The following summarizes this process:

IF…	THEN…
an institution wants to use the *Praxis II* tests in a state that has not authorized the content assessment for licensure **AND** the state requires use of a *Praxis II* content test for licensure	that institution should review test specifications to confirm that the content is necessary for student teaching and that students were exposed to the curriculum that covers the appropriate content.

Entrance into Graduate-level Teacher Programs. Graduate-level teacher programs most often focus on providing additional or advanced pedagogical skills. These programs do not typically focus on content knowledge itself. Because of this, such programs expect students to enter with sufficient levels of content knowledge. In states that use *Praxis II* content assessments for licensure, sufficient content knowledge may be defined as the candidate'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 authorized the use of the test content, that program should review the test specifications to confirm that the content is a necessary prerequisite for entrance into the program. The program also must establish a minimum score for entrance and is encouraged to use other student qualifications, in addition to the test scores, when making final entrance decisions.

Furthermore, the test should not be used to rank candidates 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 their needs, ETS:

- Schedules and implements a standard-setting study
- Helps the state analyze the results of the study
- Works with the state to help it establish the passing score.

Standard-Setting Studies

A standard-setting study produces a passing-score recommendation. A passing score is the minimum test score that a test taker needs to pass the particular licensure test and be awarded a license to teach. Each state sets its own passing score. *ETS does not set passing scores*; that is the licensing agencies' responsibility.

Standard-setting studies serve two purposes. First, they are designed to identify the level of knowledge for a teacher candidate to be considered minimally qualified for independent, beginning practice. The level of knowledge is represented by a minimum test score that candidates need to achieve. Second, the studies are designed to reconfirm the relevance (validity) of the test content for teachers in the adopting state.

Different standard-setting approaches are used for different test structures. In other words, there is a preferred standard-setting method for MC test items and another for CR test items. ETS recommends and implements a modified Angoff method for MC items and a Benchmark method for CR items. One or more ETS standard-setting specialists conduct and facilitate each standard-setting study. For each study, a technical report is produced that describes the selection and

"representativeness" of the participants involved and summarizes the standard-setting methods and results.

Panel Formation

For each method, the state (licensing agency) selects a panel of teachers and teacher educators to serve on the standard-setting panel. ETS works closely with the licensing agency to identify the appropriate types and numbers of educators from the state. ETS supplies the licensing agency with written descriptions of recommended qualifications and demographic characteristics of educators. Panels typically consist of 10 to 15 persons, the majority of whom are practicing, licensed teachers in the content area covered by the test; teacher educators, who prepare teacher-candidates, are often represented. States are encouraged to select a panel of educators that reflects the diversity in the state (e.g., racial/ethnic, gender, geographic, setting).

ETS reviews the nominations and identifies those panelists who meet the criteria. The state licensing agency is then asked to confirm and approve the panel composition. ETS convenes the panel and conducts the study using the method suitable for the type of test being reviewed by the panel.

Angoff Method of Standard Setting

This method is used for MC test items, which each have a single correct answer. In brief, this method necessitates that each panelist review each test item and judge the percentage of a hypothetical group of 100 minimally qualified test takers who would answer the item correctly. For each item, panelists record the percentage (e.g., 10%, 20%, . . . 90%) of the 100 hypothetical test takers who they feel would answer the item correctly. The judgments for each panelist (across items) are added, and the average across panelists is computed. This average represents the passing score study value. Before rendering their item judgments, panelists take the test and self-score it; define the knowledge and skills of minimally qualified test takers; receive appropriate training; and practice making standard-setting judgments. Panelists also are asked to verify that the test content is valid for use in that state.

Benchmark Method of Standard Setting

ETS uses the Benchmark method for items that require constructed responses. In this method, each panelist reviews the item, scoring rubric, and examples of candidates' performances that are clearly illustrative of the scale points on the scoring rubric for an item. Panelists are then asked to identify which benchmark performance (scale point) is most likely to be earned by a minimally qualified test taker. If a test consists of both MC and CR items, the recommended number of points from the MC section is combined with the points from the CR section to arrive at the recommended test-level passing score. For an all-CR test, the passing score is the average number of points recommended by the panel.

Before rendering their item judgments, panelists respond to the CR items and self-score them, define the knowledge and skills of just-qualified test takers, and receive appropriate training and practice making standard-setting judgments. Panelists also are asked to verify that the test content is valid for use in that state.

Standard-Setting Reports

Approximately six weeks after the standard-setting study is completed, the state receives a study report documenting who participated, the procedures and methods used, and the results. The report also includes information about the standard error of the test and passing score recommendations within one and two standard errors of the panel's recommendation. States may use this data and other state-specific information to decide on the operational passing score.

Psychometric Properties

Introduction

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

The psychometric procedures explained in this section follow well-established, relevant standards in *Standards for Educational and Psychological Testing* (1999) and the *ETS Standards for Quality and Fairness* (2002). They are used extensively in the *Praxis* program and are accepted by the psychometric community at large.

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

Test-Scoring Process

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

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

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

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

  Items that show very high DIF statistics are reviewed by a fairness panel of content experts, which often include representatives of the subgroups used in the analysis. The fairness panel decides if a test takers' performance on any item is influenced by factors not related to the construct being measured by the test. Such items are then excluded from the test scoring. A more detailed account of the DIF procedures followed by the *Praxis* program are provided in

"Differential Item Functioning (DIF) Analyses" on page 33,and are described at length in Holland and Wainer's (1993) text.

- Test developers consult with content experts or content advisory committees to determine whether all items in new test forms meet ETS's standards for quality and fairness. Their consultations are completed within days after the administration of the test.

- Statistical equating and scaling is performed on each new test approximately three weeks after the test is administered.

- Scores are sent to test takers and institutions of higher education four weeks after the test administration.

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

Item Analyses

Classical Item Analyses

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

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

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

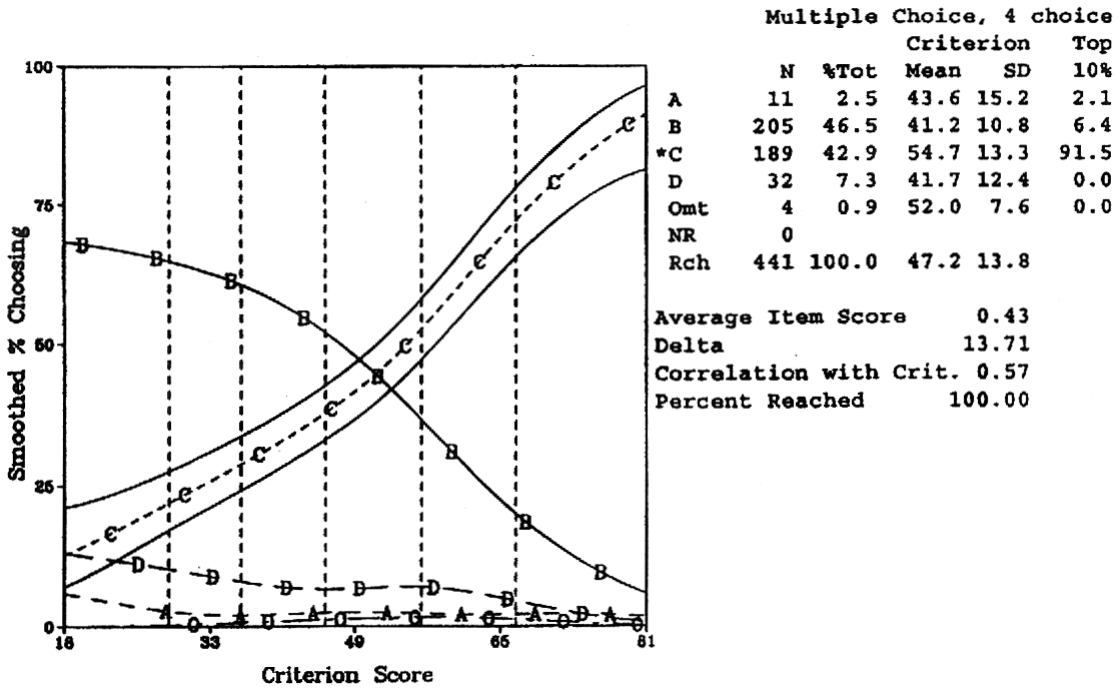

Figure 2. Example of an item analysis graph for an MC item

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

- For each response option, the table shows the count and percent of test takers who chose the option, the criterion score mean and standard deviation of respondents, and the percent of respondents with scores in the top ten percent of test takers who chose the option. The specified percentage of top scores may differ from ten percent, depending on factors such as the nature of the test and sample size.
- Four statistics are presented for the item as a whole: 1) The Average Item Score (the percent of correct responses to an item that has no penalty for guessing); 2) Delta, an index of item difficulty that has a mean of 13 and standard deviation of 4 (see footnote on page 32); 3) The correlation of the item score with the criterion score. (For an MC item this is a biserial correlation, a measure of correspondence between a normally distributed continuous variable assumed to underlie the dichotomous item's outcomes, and the criterion score); 4) the percent of test takers who reached the test item.

For CR items, both item and scorer analyses are conducted. The item analyses include distributions of scores on the item; two-way tables of rater scores before adjudication of differences between scorers; the percentage of exact and adjacent agreement; the distributions of the adjudicated scores;

and the correlation between the scores awarded by each of the two scorers. For each scorer, his/her scores on each item are compared to those of all other scorers for the same set of responses.

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

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

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

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

Speededness

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

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

All four of these indices need not be met for a test to be considered speeded. If the statistics show that many test takers did not reach several of the items, this information can be interpreted as strong evidence that the test (or a section of a test) was speeded. However, even if all or nearly all of the test takers reached all or nearly all of the items, it would be wrong to conclude, without additional information, that the test (or section) was unspeeded. Some test takers might well have answered more of the items correctly if given more time. Item statistics, such as the percent correct and the item total correlation, may help to determine whether many test takers are guessing, but the statistics

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

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

could indicate that the items at the end of the test are difficult. A *Praxis I* or *Praxis II* test will be considered speeded if more than one of the speededness indices is exceeded.

Differential Item Functioning (DIF) Analyses

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

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

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

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

DIF Statistics

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

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

The Hispanic subgroup comprises test takers who coded:

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

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

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

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

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

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

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

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

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

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

Test-Form Equating

Overview

Each *Praxis* test comprises multiple test forms, with each containing a unique set of test questions, whether multiple choice, constructed response, or a combination of both. *ETS Standards for Quality and Fairness* (2002) require the use of equating methodologies when "results … on different forms of an assessment are to be treated as though they were equivalent" (page 45), as is the case for all *Praxis* tests. Equating adjusts scores on different test forms to account for the inherent inability to produce test forms with identical degrees of difficulty, even when test-assembly processes are tight. Because equating adjusts for differences in difficulty across different *Praxis* test forms, a given scale score represents the same level of achievement for all forms of the test. Well-designed equating procedures maintain the comparability of scores for a test and thus avoid penalizing test takers who happen to encounter a selection of questions that proves to be more difficult than expected (von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004).

Scaling

To avoid confusion between the adjusted and unadjusted scores, the *Praxis* program has typically reported the adjusted scores on a score scale that makes them clearly different from the unadjusted (raw) scores. This score scale is a mathematical conversion (or scaling) of the raw scores into scaled scores with predetermined lower and upper limits. Most *Praxis* tests use a scaled score range of 100 to 200 for score reporting, although a small number of test titles use an older 250 to 990 score scale. The three subject areas of the *Praxis I* tests each have a score range of 150 to 190. The use of a scale common to all forms of the same test title enables the users of the test to compare scores on test forms that may differ slightly in difficulty.
When the first form of a *Praxis* test consisting only of MC items is administered for the first time, the method used to establish the reported score scale is as follows:

1. The raw score to be expected by guessing randomly at each item = C
   where C = Test Length * (1 / number of MC options).

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

2. The score T is defined as: Test Length * .95

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

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

   A = (Scale Maximum – Scale Minimum) / (T – C), and

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

Equating

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

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

The NEAT Design

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

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

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

2. Tests containing CR items only are equated using an internal anchor test (comprising CR items only) if the number of items in the total test is six or more.
3. Tests containing fewer than six CR items are equated using an external anchor comprising MC items that measure comparable skills and knowledge. For example, the *Praxis* English Language, Literature, and Composition: Essays test is a CR test of four essay questions. Test takers often complete the *Praxis* English Language and Literature: Content Knowledge (MC) test at the same test administration. Because the constructs measured by both tests are similar, the MC scores are used as an external anchor test. A type of external anchoring also is used for Trend Scored tests (see Scoring Methodology).
4. Tests containing sufficient numbers of both MC and CR items are equated using a combination of MC and CR items as an internal anchor test.
5. Tests containing MC items and a small number of CR items are equated using only the MC items in an internal anchor test.

The Equivalent Groups Design

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

The Single Group Design

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

The Testlet Design

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

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

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

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

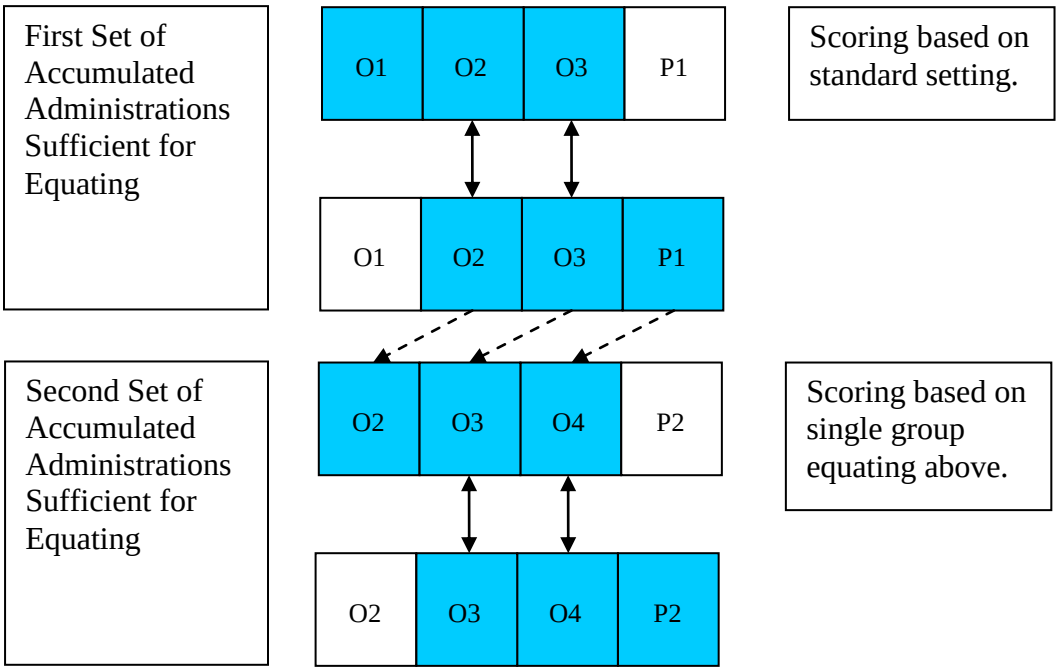

Figure 3. The Testlet Equating Design

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

Equating Methodology Summary

Because the equivalent groups equating design requires a large volume of test takers to produce dependable results, only the *Praxis I* tests use this method. Thirteen of the smallest volume *Praxis II* tests use the testlet design. All other *Praxis* tests use the NEAT design to equate new test forms.

Test Statistics

Reliability

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

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

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

If the test is not highly speeded, the KR 20 reliability estimate will be an adequate estimate of alternate-form reliability. However, because *Praxis* tests are used to make pass/fail decisions, information about the reliability of classification (RELCLASS) also is relevant to the issue of test reliability. RELCLASS is described in more detail on page 40.

Standard Error of Measurement

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

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

The standard error of measurement for the scaled score is:

$$SEM_S = A^* SEM_X \quad .$$

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

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

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

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

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

Reliability of Classification

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

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

Reliability of Scoring

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

1. The inter-rater correlations for each item are obtained from the two independent ratings, and the inter-rater reliabilities are computed from them using the Spearman-Brown formula.

2. Variance errors of scoring for each item are calculated by multiplying the item's variance by $(1 - r_{cis})$, where r_{cis} is the item's inter-rater reliability.

3. The variance errors of scoring for all of the items are added together to form the variance of errors of scoring for the entire test.

4. The standard error of scoring is defined as the square root of the variance errors of scoring for the sum obtained in step 3.

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

Scoring Methodology

Scoring

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

For tests consisting only of CR items, raw scores are a weighted composite of scores on individual items. For each question, the written responses are read and scored by two qualified scorers who are trained to score the responses to that item according to a pre-specified scoring rubric[7]. The ratings that the scorers assign are based on a rubric developed by educators who are specialists in the subject area. All scorers receive training before they score operational responses. The score on any single CR test item is the sum of the scores for CR items as assigned by the two scorers.

For tests that include both MC and CR questions, raw scores are a weighted composite of the raw MC score and the scores on the individual CR items. A test taker's score in the MC 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 *Tests at a Glance* documents, at www.ets.org/praxis.

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 MC items. The time suggested for a response can vary from 10 minutes to 60 minutes. Scoring can be:

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

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

Every test that contains CR items has a General Scoring Guide (GSG), which is written to verify that well-trained, calibrated scorers will be able to consistently evaluate responses according to clearly specified indicators. Question-specific scoring guides (QSSG) and scoring notes also are developed to inform scorers of some of the item-specific features that a response might contain. Final ratings are assigned to a response after a careful reading to find the evidence that the item has been answered. That evidence then is evaluated by selecting the set of descriptors in the scoring guide that best fits the

[8] 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 papers.

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 finalized at the "sample pulling" before the first scoring of a prompt. Sample pulling 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 papers
- Document the rationales for the agreed-upon scores
- Arrange the selected papers 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 candidate receives a fair and appropriate score
- That all candidates 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 papers: exemplars of each score point on the score scale, usually at the mid-range of a score point
- Training papers: 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 sample papers received the given score, providing consistency in what is said during training
- Calibration papers: 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. These papers are used before live scoring. Scorers are said to be calibrated when their individual ratings on a set of common CR responses are consistent with scores assigned by other scorers (known also as the "set score"). If a scorer's scores are not consistent with the set score, then she/he is required to

be retrained. Calibration verifies to some degree that ratings assigned to a given CR response by different scorers within and between different testing administrations are not very discrepant.

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

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

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

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

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

Consistency in the scoring of a form is verified by:

- Training notes that clearly indicate how an item should be interpreted
- Annotations/evidence sheets that clearly indicate how individual papers 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 the possible impact of bias that scorers may bring to the scoring session

- Calibration of scorers to ensure that they perform the scoring consistently from administration to administration

Content Category Information

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

On a test consisting only of MC items, "raw points" means the number of items answered correctly. On test consisting only of CR items, "raw points" means the sum of the ratings that the scorers awarded to the answer.

ETS provides institutions of higher education (IHEs) with the same level of individual student category information that the company provides to test takers because of IHEs' 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 IHEs to consider other information about a student's understanding in addition to category scores when making instructional decisions for students.

Quality Assurance Measures

MC 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 MC items on computer-delivered tests are automatically verified before scores are reported.)

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

Appropriate Score Use

ETS is committed to furthering quality and equity in education by providing valid and fair tests, research, and related services. Central to this objective is helping those who use the *Praxis* tests to understand what are considered their proper uses. The booklet *Proper Use of The Praxis Series and*

Related Assessments defines proper test use as adequate evidence to support the intended use of the test and to support the decisions and outcomes rendered on the basis of test scores.

Proper assessment use is a joint responsibility of ETS as the test developer, and of states, agencies, associations, and institutions of higher education as the test users. The *Praxis* program is responsible for developing valid and fair assessments in accordance with technical guidelines established by the American Educational Research Association, the American Psychological Association, and the National Council on Educational Measurement in Education (*Standards for Educational and Psychological Testing*, 1999).

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

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

Score Reporting

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

Scanning/Scoring

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

Score Reports

Each test taker receives a detailed score report that includes the test taker's overall score, passing status and, if applicable, information regarding performance on specific areas of the test. The report also includes explanatory materials to help the candidate 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 and Hearing Association (ASHA)
- The National Association of School Psychologists℠ (NASP℠)
- Department of Defense Dependent Schools (DODDS)
- Any other entity designated to receive scores by the state or law.

Score Information for States and Institutions

When paper 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.

States, agencies, and institutions choose how they want score reports delivered: via electronic download, CD-ROM, paper report, or pressure-sensitive labels.

A State Agency Summary Report is for state departments of education. It includes:

1. The institutions within that state whose students took the *Praxis* tests.
2. Frequency distributions of total test scaled scores for the state as a whole and separately by educational institutions.
3. Demographic performance breakdowns of the test-taking population.
4. Frequency distributions of test scores broken down by gender, ethnicity, educational level, undergraduate and graduate majors.

An Institutional Summary Report is for institutions of higher education. It includes:

1. The number of individuals from the institution who took a *Praxis* test.
2. Demographic breakdowns of the test takers by gender, ethnicity, and undergraduate and graduate majors.
3. Frequency distributions and summary statistics of scaled scores for both the national and institutional samples of test takers for each *Praxis* test.
4. Separate frequency distributions of scaled scores and summary statistics for each demographic group.
5. Summary statistics comparing the performance of students at the school with statewide and national samples (if the test items can be divided into separate categories).

View a sample Institutional Summary Report.

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 program completer data for each teacher preparation program.

This database system:

- Collects program completer data from each teacher preparation program
- Matches each completer's information with the correct test by licensure area
- Lists all program completers 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 and attends state workshops 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 program completer data, schedules for relevant due dates, and statistical support in interpreting the passing-rate data.

ETS also maintains a telephone hotline and e-mail service 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 — *Praxis* Job Analyses

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

Table 1 – *Praxis*-Related Job Analyses

Job Analysis
Audiology
Biology
Business Education
Chemistry and Physics
Deaf and Hard of Hearing
Early Childhood: Content Knowledge
Earth and Space Science
Elementary School Teachers
Elementary School Survey of CA Teachers
English
French
Fundamental Subjects: Content Knowledge (FS:CK)
General Principles of Teaching and Learning
General Science
German
Guiding Conception and Assessment Principles for *The Praxis Series*
Knowledge at Elementary Level
Mathematics
Middle School Teachers
Middle School General Science
Middle School Language Arts
Middle School Mathematics
Middle School Social Studies
Multi-State
Music
Praxis I
ParaProfessional Assessment
Physical Education
School Administrator
School Psychologist
School Social Worker
School Superintendent
Secondary School Teachers
Social Studies
Spanish
Special Education

Appendix B – Statistical Characteristics of *Praxis I* and *Praxis II* Tests

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

- **Range** — The lowest to the highest scaled score possible on any edition of the test. The actual maximum and minimum possible scores for a given form of a test may differ from one edition of a test to another.

- **Interval** — The number of points separating the possible score levels. If the score interval is 10, for example, only scores divisible by 10 are possible.

- **Number of Test Takers** — The number of people taking the test within the time period listed in the notes following the table.

- **Median** — The score that separates the lower half of the scores from the upper half, calculated for the scores obtained by the group of test takers listed in the notes following the table.

- **Average Performance Range** — The range of scores earned by the middle 50 percent of the test takers, calculated for the group of test takers listed in the notes following the table. This range provides an indication of the difficulty of the test.

- **Mean** — The arithmetic average, calculated for the scores obtained by the group of test takers listed in the notes following Table 1.

- **Standard Deviation** — The amount of variability among the scores obtained by the group of test takers listed in the notes following Table 1.

- **Standard Error of Measurement** — The standard error of measurement (SEM) is a test statistic described on page 39 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 an examinee's scores are from one testing to another, the smaller the SEM. The SEM is included in Table 1 for *The Praxis Series* tests that have at least six items. Because estimates of the standard error may vary slightly from one test administration to another and from one test edition to another, the tabled values are averages of the SEMs obtained from all forms of the test currently in use.

- **Standard Error of Scoring** — For tests in which the scoring involves human judgment, this statistic describes the reliability of the process of scoring the test takers' responses. A test taker's score on one of these tests will depend to some extent on the particular scorers who rate her/his responses. The more consistent the ratings assigned to the same responses by different scorers,

the smaller the standard error of scoring (SES). If a large number of test takers take a test for which the standard error of scoring is four points, about two-thirds of them will receive scores within four points of the scores that they would get if their responses were scored by all possible scorers. The SES is included in Table 1 for tests in *The Praxis Series* assessments consisting of CR items. The tabled values are averages of the SESs obtained from all forms of the test currently in use. Since the January 2008 *Praxis* test administration, all CR tests have been scored by two independent raters. The standard error of scoring for a test consisting only of MC items is zero, because MC scoring is a purely mechanical process with no possibility of disagreement between scorers.

- **Reliability** — The reliability coefficient is an estimate of the correlation between examinees' test scores and the scores they might have achieved on different forms of the same test. Its value ranges from zero to one. For all *Praxis* tests with at least six items, this index is calculated using an internal consistency estimate (Kuder and Richardson, 1937), based on the statistical relationships among the test takers' responses to all items in the test. The reliability of a test may vary slightly from one test administration to another and from one form of the test to another. The tabled values are averages of the reliabilities obtained from all the forms of the test currently in use.

Table 1 — Statistical Characteristics of *Praxis I* and *Praxis II* 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 (0700)	250 – 990	10	910	570	530 – 610	570.9	66.0	29	0	0.87
Agriculture (PA) (0780)	250 – 990	10	70	660	610 – 710	660.8	72.3	33	0	0.84
Agriculture (OR) (0900)	250 – 990	10	35	750	710 – 770	729.4	75.7	30	0	0.87
Art Making (0131)	100 – 200	1	2126	162	155 – 169	161.1	13.2	q	5.2	0.87
Art: Content Knowledge (0133)	100 – 200	1	10195	171	162 – 180	170.9	12.9	4.7	0	0.88
Art: Content, Traditions, Criticism, and Aesthetics (0132)	100 – 200	5	1661	145	145 – 155	147.6	10.5	q	4.5	0.85
Audiology (0340)	250 – 990	10	1466	640	620 – 670	644.3	30.0	13	0	0.87
Biology and General Science (0030)	250 – 990	10	1141	650	580 – 700	637.8	77.8	19	0	0.94
Biology: Content Essays (0233)	100 – 200	1	725	150	145 – 158	150.9	12.4	q	3.2	0.94
Biology: Content Knowledge (CT) (0235)	100 – 200	1	13735	162	152 – 173	161.2	17.1	4.5	0	0.94
Biology: Content Knowledge, Part 1 (0231)	100 – 200	1	523	168	156 – 179	167.0	17.2	6.1	0	0.89
Braille Proficiency (0631)	100 – 200	1	f	f	f	f	f	f	f	f
Business Education (0100)	250 – 990	10	9009	650	610 – 680	648.1	53.2	19	0	0.90
Business Education: Content Knowledge (0101)	100 – 200	1	f	f	f	f	f	f	f	f
Chemistry, Physics, and General Science (0070)	250 – 990	10	599	580	520 – 650	580.7	89.2	26	0	0.92
Chemistry: Content Essays (0242)	100 – 200	5	270	155	145 – 165	156.6	16.1	q	4.3	0.93
Chemistry: Content Knowledge (CT) (0245)	100 – 200	1	5410	160	148 – 174	159.7	20.3	6.3	0	0.90
Citizenship Education: Content Knowledge (0087)	100 – 200	1	388	165	155 – 176	165.5	14.4	5	0	0.89

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Communication (0800)	250 – 990	10	344	725	660 – 780	719.3	88.3	35	0	0.81
Computerized PPST® – Math (5730)	150 – 190	1	98594	179	174 – 183	177.9	6.2	2.5	0	0.88
Computerized PPST – Reading (5710)	150 – 190	1	101181	178	174 – 181	176.8	5.7	2.4	0	0.87
Computerized PPST – Writing (5720)	150 – 190	1	98395	175	173 – 178	175.4	4.2	2.6	0.3	0.68
Cooperative Education (0810)	250 – 990	10	106	820	780 – 850	810.5	52.5	29	0	0.72
Driver Education (0867)	100 – 200	1	120	174	168 – 182	174.1	9.3	5	0	0.76
Early Childhood Education (0020)	250 – 990	10	4205	650	600 – 690	637.6	70.7	23	0	0.89
Early Childhood: Content Knowledge (0022)	100 – 200	1	11481	176	168 – 183	174.7	11.7	4.4	0	0.89
Earth and Space Sciences: Content Knowledge (0571)	100 – 200	1	3605	163	151 – 176	161.9	18.7	5.4	0	0.92
Economics (0910)	250 – 990	10	437	560	500 – 640	569.6	101.5	37	0	0.85
Education of Deaf and Hard of Hearing Students (0271)	100 – 200	1	462	172	166 – 178	171.4	9.8	6.7	3.3	0.67
Education of Exceptional Students: Core Content Knowledge (0353)	100 – 200	1	35826	174	165 – 183	173.3	13.6	7.3	0	0.77
Education of Exceptional Students: Learning Disabilities (0382)	100 – 200	1	510	169	160 – 178	168.7	15.2	8.7	1.5	0.68
Education of Exceptional Students: Mild to Moderate Disabilities (0542)	100 – 200	1	13461	179	171 – 186	178.1	11.1	q	2.6	0.95
Education of Exceptional Students: Severe to Profound Disabilities (0544)	100 – 200	1	1595	166	158 – 177	167.9	15.5	q	3.6	0.96
Education of Young Children (0021)	100 – 200	1	13580	184	176 – 190	182.5	10.9	6.3	2.5	0.70
Educational Leadership: Administration and Supervision (0410)	250 – 990	10	15721	710	650 – 750	699.1	71.7	26	0	0.88
Educational Leadership: Administration and Supervision (0411)	100 – 200	1	f	f	f	f	f	f	f	f
Elementary Education: Content Area Exercises (0012)	100 – 200	1	24149	157	151 – 163	157.0	9.0	q	4.1	0.85
Elementary Education: Content Knowledge (0014)	100 – 200	1	95773	164	152 – 177	163.5	17.3	5.7	0	0.91
Elementary Education: Content Knowledge (5014)	100 – 200	1	f	f	f	f	f	f	f	f
Elementary Education: Curriculum, Instruction, and Assessment (0011)	100 – 200	1	67129	177	168 – 185	174.9	14.0	6.6	0	0.85
Elementary Education: Curriculum, Instruction, and Assessment (5011)	100 – 200	1	f	f	f	f	f	f	f	f
English Language, Literature, and Comp.: Content Knowledge (0041)	100 – 200	1	36180	177	166 – 188	175.8	15.5	4.7	0	0.92

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
English Language, Literature, and Comp.: Essays (0042)	100 – 200	5	5124	160	150 – 165	158.4	10.5	q	3.5	0.91
English Language, Literature, and Comp.: Pedagogy (0043)	100 – 200	5	5919	155	145 – 165	154.8	13.8	q	4	0.91
English to Speakers of Other Languages (0360)	250 – 990	10	9401	640	570 – 700	640.4	98.9	33	0	0.91
English to Speakers of Other Languages (0361)	100 – 200	1	f	f	f	f	f	f	f	f
Environmental Education (0830)	250 – 990	10	242	710	650 – 770	701.1	100.3	39	0	0.86
Family and Consumer Sciences (0121)	100 – 200	1	2297	170	163 – 177	169.3	12.6	4.6	0	0.89
French: Content Knowledge (0173)	100 – 200	1	2038	181	166 – 193	178.0	17.5	4.4	0	0.95
French: World Languages (5174)	100 – 200	1	f	f	f	f	f	f	f	f
Fundamental Subjects: Content Knowledge (0511)	100 – 200	1	26467	175	163 – 185	173.5	15.0	5.4	0	0.89
General Science: Content Essays (0433)	100 – 200	5	661	150	130 – 155	145.2	18.1	q	3.1	0.96
General Science: Content Knowledge (0435)	100 – 200	1	8703	166	153 – 179	165.1	18.9	5.5	0	0.92
General Science: Content Knowledge, Part 1 (0431)	100 – 200	1	971	164	152 – 175	163.0	16.9	6.7	0	0.86
General Science: Content Knowledge, Part 2 (0432)	100 – 200	1	540	161	151 – 173	160.6	16.6	7.1	0	0.84
Geography (0920)	250 – 990	10	813	680	620 – 740	667.6	93.9	31	0	0.90
German: Content Knowledge (0181)	100 – 200	1	691	183	167 – 195	178.3	20.2	4.7	0	0.95
German: World Languages (5183)	100 – 200	1	f	f	f	f	f	f	f	f
Gifted Education (0357)	100 - 200	1	287	161	156 - 167	160.8	10.4	5.4	0	0.76
Government/Political Science (0930)	250 - 990	10	898	690	620 - 760	680.3	97.5	28	0	0.92
Health and Physical Education: Content Knowledge (0856)	100 - 200	1	6081	164	156 - 170	162.7	11.2	4.9	0	0.84
Health Education (0550)	250 - 990	10	6479	700	650 - 740	695.2	64.4	28	0	0.82
Interdisciplinary Early Childhood Education (0023)	100 - 200	1	321	178	169 - 184	174.7	13.1	5.6	0	0.80
Introduction to the Teaching of Reading (0200)	250 - 990	10	5157	640	600 - 690	638.4	61.9	24	0	0.87
Latin (0600)	250 - 990	10	209	730	650 - 850	735.0	133.8	35	0	0.94
Library Media Specialist (0311)	100 - 200	1	1594	166	156 - 174	164.3	13.8	5.3	0	0.89
Life Science: Pedagogy (0234)	100 - 200	1	1470	156	149 - 159	154.5	8.9	q	2	0.95
Marketing Education (0561)	100 - 200	1	760	172	162 - 181	170.5	14.8	5.6	0	0.86
Mathematics: Content Knowledge (0061)	100 - 200	1	26350	145	128 - 160	143.8	22.8	7.5	0	0.89
Mathematics: Pedagogy (0065)	100 - 200	5	3501	145	130 - 160	145.3	21.8	q	5.4	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
Mathematics: Proofs, Models, and Problems, Part 1 (0063)	100 - 200	1	2972	163	150 - 179	164.5	18.5	q	3.7	0.97
Middle School English Language Arts (0049)	100 - 200	1	17930	174	163 - 184	172.2	16.8	7.1	2.2	0.84
Middle School Mathematics (0069)	100 - 200	1	28485	163	152 - 176	163.7	17.9	7.1	1.2	0.84
Middle School Science (0439)	100 - 200	1	13035	158	147 - 171	159.2	16.9	6.1	2.3	0.88
Middle School Social Studies (0089)	100 - 200	1	12722	166	155 - 179	166.3	17.5	6.1	1.9	0.89
Middle School: Content Knowledge (0146)	100 - 200	1	10070	162	151 - 174	161.8	16.4	5.6	0	0.90
Music: Analysis (0112)	100 - 200	1	1115	173	164 - 179	170.9	12.1	q	0.8	0.99
Music: Concepts and Processes (0111)	100 - 200	5	2651	155	145 - 165	157.1	13.8	q	1.9	0.99
Music: Content Knowledge (0113)	100 - 200	1	11091	167	158 - 176	166.5	12.9	5.3	0	0.86
ParaPro Assessment (0755)	420 - 480	1	18356	470	460 - 477	466.5	12.3	3.4	0	0.94
ParaPro Assessment (1755)	420 - 480	1	50721	470	462 - 477	467.6	11.1	3.2	0	0.94
Physical Education: Content Knowledge (0091)	100 - 200	1	14524	156	150 - 162	155.3	9.5	4.1	0	0.83
Physical Education: Movement Forms - Analysis/Design (0092)	100 - 200	1	2915	161	157 - 165	160.8	6.2	q	2.1	0.92
Physical Education: Movement Forms - Video Evaluation (0093)	100 - 200	5	1406	165	160 - 170	164.3	9.3	q	4.6	0.86
Physical Science: Content Knowledge (0481)	100 - 200	1	1048	164	152 - 178	163.9	17.3	6.3	0	0.88
Physical Science: Pedagogy (0483)	100 - 200	1	862	163	154 - 171	162.5	13.8	q	4.5	0.92
Physics: Content Essays (0262)	100 - 200	5	145	165	150 - 180	164.8	19.6	q	7.8	0.83
Physics: Content Knowledge (0261)	100 - 200	1	193	135	116 - 153	135.0	23.0	7.4	0	0.88
Physics: Content Knowledge (CT) (0265)	100 - 200	1	2828	150	134 - 166	149.4	22.4	6.3	0	0.92
Pre-Kindergarten Education (0530)	250 - 990	10	330	690	650 - 750	687.5	74.4	29	0	0.82
PPST: Mathematics (0730)	150 - 190	1	51539	179	173 - 183	178.0	6.9	2.8	0	0.87
PPST: Reading (0710)	150 - 190	1	52367	178	174 - 181	177.1	5.5	2.3	0	0.87
PPST: Writing (0720)	150 - 190	1	51924	176	173 - 178	175.5	3.9	2.3	0.3	0.72
Principles of Learning and Teaching: Early Childhood (0521)	100 - 200	1	14618	184	176 - 191	182.5	11.8	7.1	1.8	0.67
Principles of Learning and Teaching: Grades 5-9 (0523)	100 - 200	1	13069	172	165 - 178	170.4	12.2	7.3	1.9	0.69
Principles of Learning and Teaching: Grades 7-12 (0524)	100 - 200	1	45638	173	167 - 180	173.1	10.7	6.2	1.5	0.72

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Principles of Learning and Teaching: Grades K-6 (0522)	100 - 200	1	46468	175	168 - 182	174.1	11.6	7.4	1.7	0.69
Psychology (0390)	250 - 990	10	321	670	600 - 760	674.0	111.3	35	0	0.90
Reading Across the Curriculum: Elementary (0201)	100 - 200	1	7731	168	160 - 176	167.8	11.6	6.6	3	0.77
Reading for Virginia Educators: Elementary and Special Education (0306)	100 - 200	1	f	f	f	f	f	f	f	f
Reading for Virginia Educators: Elementary and Special Education (5306)	100 - 200	1	f	f	f	f	f	f	f	f
Reading for Virginia Educators: Reading Specialist (0304)	100 - 200	1	f	f	f	f	f	f	f	f
Reading for Virginia Educators: Reading Specialist (5304)	100 - 200	1	f	f	f	f	f	f	f	f
Reading Specialist (0300)	250 - 990	10	14061	580	520 - 630	573.0	73.4	27	0	0.85
Safety/Driver Education (0860)	250 - 990	10	236	570	520 - 610	567.3	74.1	41	0	0.69
School Guidance and Counseling (0420)	250 - 990	10	11155	660	620 - 700	654.6	61.4	24	0	0.89
School Leaders Licensure Assessment (1011)	100 - 200	1	7268	173	166 – 179	171.8	10.4	5.3	2.3	.76
School Superintendents Assessment (1020)	100 - 200	1	1765	170	164 – 176	170.1	8.9	5.3	1.6	.54
School Psychologist (0401)	100 - 200	1	4771	174	168 - 180	173.3	8.9	3.6	0	0.87
School Social Worker: Content Knowledge (0211)	100 - 200	1	211	179	174 - 184	178.4	9.1	5.7	0	0.73
Social Studies: Analytical Essays (0082)	100 - 200	5	996	150	145 - 155	149.7	10.3	q	4.2	0.89
Social Studies: Content Knowledge (0081)	100 - 200	1	31654	167	157 - 177	166.6	15.0	4.9	0	0.91
Social Sciences: Content Knowledge (0951)	100 - 200	1	286	156	145 - 166	155.8	15.0	5.5	0	0.86
Social Studies: Interpretation of Materials (0083)	100 - 200	1	2942	165	160 - 172	165.3	10.3	q	2.9	0.93
Social Studies: Pedagogy (0084)	100 - 200	1	2348	173	164 - 184	172.2	16.4	q	3.3	0.97
Sociology (0950)	250 - 990	10	128	670	630 - 730	673.4	86.4	32	0	0.89
Spanish: Content Knowledge (0191)	100 - 200	1	10209	176	164 - 188	174.0	17.4	4.8	0	0.94
Spanish: World Languages (5195)	100 - 200	1	f	f	f	f	f	f	f	f
Special Education: Application of Core Principles Across Categories of Disability (0352)	100 - 200	1	7144	149	141 - 158	149.3	13.4	7.7	0	0.69

Test	Scale Range	Interval	No. of Test Takers	Median	Average Performance Range	Mean	Standard Deviation	Standard Error of Measurement	Standard Error of Scoring	Reliability
Special Education: Core Content Knowledge and Applications (0354)	100 - 200	1	f	f	f	f	f	f	f	f
Special Education: Core Knowledge and Mild to Moderate Applications (0543)	100 - 200	1	f	f	f	f	f	f	f	f
Special Education: Core Knowledge and Severe to Profound Applications (0545)	100 - 200	1	f	f	f	f	f	f	f	f
Special Education: Knowledge-Based Core Principles (0351)	100 - 200	1	5087	160	152 - 172	159.9	15.3	8.6	0	0.75
Special Education: Preschool/Early Childhood (0690)	250 - 990	10	2068	630	590 - 670	628.5	67.8	32	0	0.80
Special Education: Teaching Students with Behavioral Disorders/Emotional Disturbances (0371)	100 - 200	1	732	162	153 - 174	161.9	16.1	9.2	0	0.67
Special Education: Teaching Students with Learning Disabilities (0381)	100 - 200	1	313	154	140 - 167	151.9	18.5	7.3	0	0.79
Special Education: Teaching Students with Mental Retardation (0321)	100 - 200	1	375	157	147 - 169	155.4	17.6	8.7	0	0.78
Speech Communication (0220)	250 - 990	10	742	670	610 - 720	663.5	77.9	28	0	0.87
Speech Communication: Content Knowledge (0221)	100 - 200	1	219	159	152 - 167	158.8	12.6	4.9	0	0.85
Speech-Language Pathology (0330)	250 - 990	10	19269	680	640 - 720	680.6	57.4	24	0	0.88
Teaching Foundations: English (0048)	100 - 200	1	128	188.5	183 - 195	187.3	11.2	6.3	1.6	0.81
Teaching Foundations: Mathematics (0068)	100 - 200	1	152	187	177 - 193	183.9	12.0	6.8	1.8	0.83
Teaching Foundations: Multiple Subjects (0528)	100 - 200	1	335	181	172 - 188	178.5	13.2	5.3	0.8	0.91
Teaching Foundations: Science (0438)	100 - 200	1	140	185	177 - 192	183.8	11.3	6.4	2.8	0.86
Teaching Reading (0204)	100 - 200	1	f	f	f	f	f	f	f	f
Teaching Speech to Students with Language Impairments (0880)	250 - 990	10	780	700	650 - 740	694.2	67.0	37	0	0.74
Teaching Students with Visual Impairments (0280)	250 - 990	10	293	750	710 - 800	748.0	70.9	32	0	0.82
Teaching Students with Visual Impairments (0281)	100 - 200	1	f	f	f	f	f	f	f	f
Technology Education (0050)	250 - 990	10	2008	650	610 - 680	645.7	50.7	16	0	0.91
Theatre (0640)	250 - 990	10	1121	700	630 - 750	686.8	84.9	35	0	0.88
Vocational General Knowledge (0890)	250 - 990	10	149	650	580 - 720	648.1	110.3	38	0	0.90
World and U.S. History: Content Knowledge (0941)	100 - 200	1	3492	159	146 - 170	158.0	16.5	5	0	0.92

Notes:

- "Number of Test Takers," "Median," and "Average Performance Range" were calculated from the records of test takers who took the test between Sept. 1, 2007 and July 31, 2010, 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 PPST® tests were calculated on college freshmen, sophomores, and juniors.
- The Median and Average Performance Range for all other tests were calculated on test takers who were college seniors, college graduates, graduate students, or holders of master's or doctoral degrees.

Legend:

q = Insufficient number of questions: SEM could not be estimated accurately for tests that include only a small number of independent questions or exercises.

f = Summary statistics are not yet available for new or rescaled tests administered for the first time in 2010-2011.

Bibliography

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological measurement*. Washington, DC: American Psychological Association.

Clauser. B.E., Margolis, M.J., & Case, S.M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational Measurement (4th Edition)* (pp. 701–731). Westport, CT: Praeger.

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

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

Educational Testing Service, *ETS Standards for Quality and Fairness*, Princeton, N.J., 2002.

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

Educational Testing Service, *Proper Use of The PRAXIS Series and Related Assessments*, Princeton, N.J., 2006

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.

Kane, M. T. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement, 2*, 135–170.

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. (2002). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues and Practice*, *21*, 25–37.

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

Schmit, 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.

Setting Performance Standards: Concepts, Methods, and Perspectives, edited by Gregory J. Cizek (2001), Mahwah, N.J.: Lawrence Erlbaum Associates.

Tannenbaum, R.J. (1999). Laying the groundwork for a licensure assessment. *Journal of Personnel Evaluation in Education, 13,* 225–244.

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.