Technical Manual for the *ETS*® Performance Assessments

June 2026

ETS® Professional Educator PROGRAMS

Table of Contents

Preface

Purpose of this manual

The purpose of this Technical Manual is to explain:

- the purpose of the *ETS*® Performance Assessments
- the approach taken by ETS in developing the *ETS* Performance Assessments
- the validity evidence supporting score use for the *ETS* Performance Assessments
- the adoption process for the *ETS* Performance Assessments
- the statistical analyses supporting the psychometric quality of *ETS* Performance Assessments
- the score reporting process
- statistical summaries of test taker performance on the *ETS* Performance Assessments

Audience

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

- knowing more about the *ETS* Performance Assessments
- how the *ETS* Performance Assessments relate to state licensure requirements
- understanding the development and scoring of the *ETS* Performance Assessments
- the statistical characteristics of the *ETS* Performance Assessments

Purpose of a Performance Assessment

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 *ETS* Performance Assessments, which provide states with assessments and ancillary services that support their educator licensure and certification process.

What is a performance assessment?

A performance assessment is a method of using authentic tasks such as activities, exercises, or problems for assessing how well test takers apply their knowledge, skills, and abilities to a particular real-world or authentic situation.

What is its purpose?

A performance assessment allows a candidate to demonstrate his or her performance during a clinical experience. Successful completion of the assessment will help to demonstrate that the candidate is prepared to begin working as an effective school leader. Since evaluators cannot be present in all schools at all times, the most authentic method of evaluation becomes the portfolio, which the candidate can submit to an objective third party. A purpose of the performance assessment is, therefore, evaluation of how well a candidate can apply their knowledge and skills.

What are the assessments?

The *ETS* Performance Assessment covered in this manual:

ETS Performance Assessment for School Leaders (PASL)

This assessment is aligned with national and state specific standards. The assessment allows the school leader candidate to demonstrate:

- the ability to address and resolve a significant problem/challenge in their school that influences instructional practice and student learning (Task 1)
- skills in establishing and supporting effective and continuous professional development with staff for the purpose of improved instruction and student learning (Task 2)
- the ability to facilitate stakeholders' efforts to build a collaborative team within the school to improve instruction, student achievement and the school culture. A 15-minute video is required with this task (Task 3)

How do the ETS Performance Assessments address the State's needs?

States have always wanted to ensure that beginning educators have the requisite knowledge and skills necessary to certify as an educator. The *ETS* Performance Assessments provide a state agency/licensing authority with appropriate tools to make decisions about applicants for educator licensure. In this way, the *ETS* Performance Assessments meet the basic licensure needs of the agency/licensing authority.

In addition to the assessments themselves, the *ETS* Performance Assessments provide the state with ancillary materials that help them make decisions related to licensure. See information to help decision makers understand how the ETS Performance Assessments support state licensure needs.

PASL Assessment: https://www.ets.org/pasl.html

States also want to ensure that their applicants' needs are being met. To that end, ETS has made available a number of helpful test preparation tools.

ETS Performance Assessment for School Leaders (PASL)

- The PASL Candidate and Educator Handbook, a valuable resource for completing the assessment. It includes a basic overview of the assessment and test-taking strategies, as well as information on:
    - how candidates should prioritize activities and build responses
    - support and ethical guidelines for task responses
    - guidelines for writing — each task requires some form of written response; it is imperative that test takers understand what kind of writing (descriptive, analytic or reflective) is required by each guiding prompt
    - collecting evidence — evidence is found in the information that test takers provide within the written commentary and in the artifacts that they submit; what they need to know about evidence, how to select evidence for tasks, and how to submit both student and teacher artifacts as evidence are described
    - how to prepare for the video recording, how to record interactions with colleagues, how to analyze a video, and the importance of practice videos; the video is meant to provide as authentic and complete a view of a test taker's instructional leadership as possible
    - scoring of tasks

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

Assessment Development

Fairness in Test Development

ETS is committed to ensuring 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 Guidelines for Developing Fair Tests and Communications.

Test Development Standards

During the test development process, ETS follows the strict guidelines detailed in *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) and:

- define clearly the purpose of the test and the claims one wants to make about the test takers
- develop test specifications and test blueprints consistent with the purpose of the test and the domains of knowledge identified as important for licensure
- develop specifications for item types and numbers of items needed to adequately sample the domains of knowledge
- develop test items that provide evidence of the measurable-behavior indicators detailed in the test specifications
- review assessments for potential fairness or bias concerns

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 (AERA, APA, & NCME, 2014).

A test developed to inform licensure[1] 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 (AERA, APA, NCME, 2014). "Licensure is designed to protect citizens from mental, physical, or economic harm that could be caused by practitioners who may not be sufficiently competent to enter the profession" (Schmitt, 1995). A licensure test is often included in the larger licensure process — which typically includes educational and experiential requirements — because it represents a standardized, uniform opportunity to determine if a test taker has acquired and can demonstrate adequate command of a domain of knowledge and/or skills that the profession has defined as being important or necessary to be considered qualified to enter the profession.

As described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), 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. 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" (AERA, APA, NCME, 2014, p. 181). A licensure test, however, should not be expected to cover all occupationally relevant knowledge and/or skills; it is only the subset of this that is most directly connected to safe and effective practice at the time of entry into the profession.

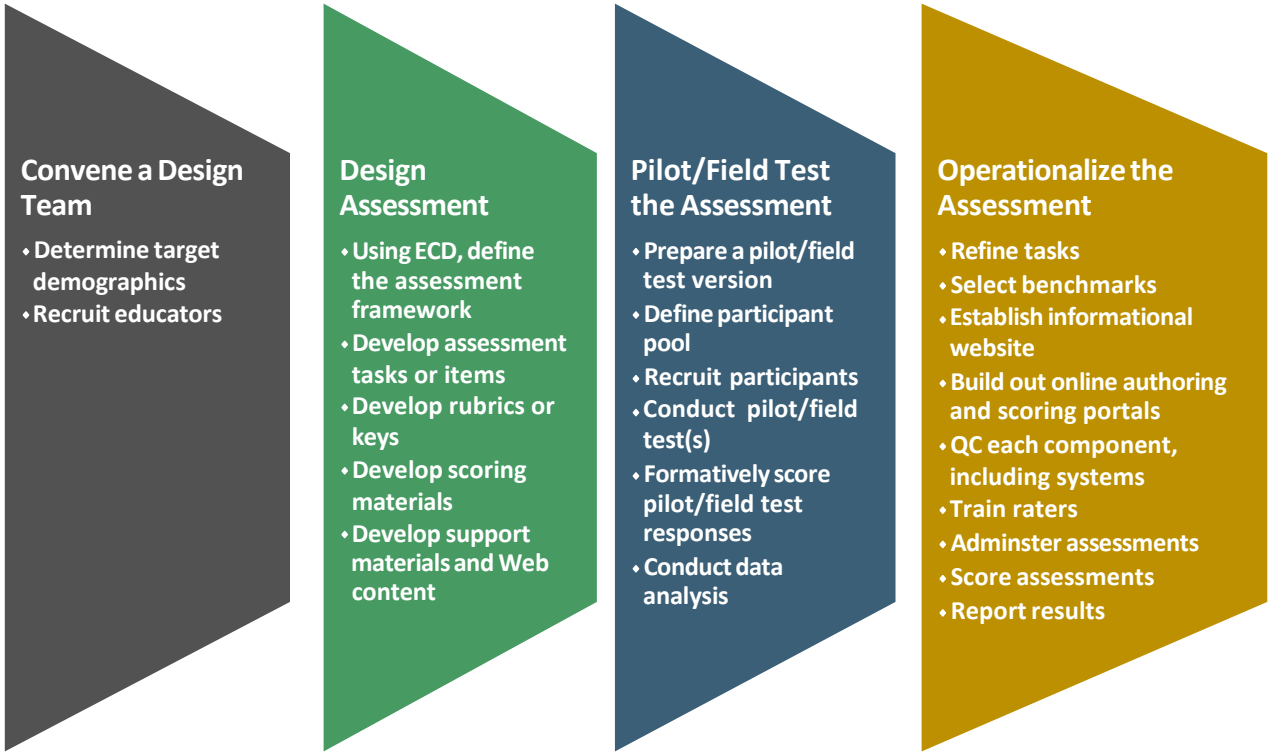
The link forged between occupational content and test content is based on expert judgment by practitioners and other stakeholders in the profession who may have an informed perspective about requisite occupational knowledge and/or skills.

Test Development Process

Evidence-centered Design (ECD)

ETS relies heavily on the evidence-centered design (ECD) process to produce high-quality, high-stakes assessments in the development of each of our assessments.[2] The following discussion outlines the process of ECD used in the design and development of the ETS Performance assessments. In developing and implementing the Standards-Based Assessments for Coursework and Clinical Experience we followed the process depicted in Figure 1.

Figure 1. Evidence-centered Design



Convening Design Teams

Our first step was to bring together design teams consisting of educators who meet specific demographic requirements. Educators and the university faculty who prepare candidates are most knowledgeable about the content and pedagogy expectations for educators at the various stages of their careers. In every step of the assessment development process educators nominated by state licensing agencies formed the core of the design teams. This is an important foundation of ECD.

Working with assessment and psychometric experts, educators make the decisions that create assessments that are true to the appropriate standards. Each partner in the relationship brings a specific set of skills and abilities that are needed for the project to succeed. In this case, ETS brings its extensive knowledge of performance assessment, validity and reliability, fairness in performance testing and scoring, and scoring unique types of assessments.

Expert educators aid the development process by bringing their in-depth knowledge of the standards and an understanding of what educators should know and be doing. Professional educators know what candidates should be able to do when they enter the classroom.

Diversity and fairness are critical to assessment design. As such, development committees should be representative of the population. In choosing participants for the development team, ETS considers the following:

- geographic location across a state or states
- developmental level taught
- content areas taught
- gender
- race/ethnicity
- years of experience

Designing Assessments – Alignment with Standards

As each new assessment was developed, the development teams "unpacked" the relevant standards as a foundation for each assessment. As we did this, we were guided by the principle that "not everything that can be measured is important and not everything that is important can be measured." As we worked with educator teams to examine the standards, we encouraged them to consider the following questions.

- Is every standard essential to measure in the performance context?
- Are there specific constructs within each standard that are most important to measure?
- How would we think about measuring the identified standards/constructs?
- What kinds of evidence could we ask educators to supply to meet this standard/construct?
- Can we design an activity, exercise, or evidence-collection device that will reasonably allow teacher candidates to provide evidence?

Who Are We Measuring?

One of the first tasks of a development committee is to achieve consensus on the characteristics of the educators who will take the assessment. It is critical that as the committee forms a consensus of these characteristics, it designs tasks that are appropriate for a specific group of candidates at the specific stage in their training or careers. We would not ask initial licensure candidates to show the same proficiency on a set of tasks as highly experienced educators. Therefore, it is of great importance that the development team understands whom the assessment is measuring, what the test takers know, and what the test takers should be able to demonstrate in regard to each of the standards.

What Claims Do We Want to Make About These Educators?

The claims should answer the question, "What do we want to say about candidates on the basis of the standards being assessed?" Claims are the statements we want to be able to make about a candidate's knowledge, skills, abilities, or other attributes on the basis of their performances on the assessment. Claims may be very general or more specific. A single assessment may be the basis for many claims at different levels of specificity. To know what we want to say about the candidate requires that we know the purpose of the assessment. Therefore, ECD begins with a clear description of the purpose of the assessment. Lower-level claims must support the high-level claim. These lower-level claims must get at the necessary knowledge, skills, and abilities at increasing levels of specificity.

What Evidence Would Support Those Claims?

Not everything that we can measure is important, and we cannot measure everything that is important. Determining what is important to measure is a critical step in the development process. We cannot measure everything that is included in the domain of practice of educators. We need to determine which things are most important to measure in light of the criteria, and what things are measurable. Further, what is measured must be appropriate for the particular population that is being assessed.

In designing the assessments, we applied the National Board for Professional Teaching Standards APPLE criteria for quality performance assessments. Such assessments must meet five criteria to be effective. They must be:

- administratively feasible
- professionally acceptable
- publicly credible
- legally defensible
- economically affordable

Evidence can take many forms. Some of the forms that evidence can take include:

- direct observation
- portfolio
- video
- student work samples/artifacts
- teacher work samples, including lesson plans, worksheets, and assessments
- published evidence, such as newspaper articles, photos, and papers
- testing documentation, such as standardized test results, teacher-created assessment data, and formative assessment data
- contextual information about classes and schools
- written explanations and documentation
- reflective pieces

Designing Tasks to Generate Evidence

A key job of the development team is to determine what — among the various constructs we could measure in order to provide evidence of a claim — is most important to measure. Then they determined what evidence would support those claims.

In thinking this through, the team members considered the following things.

- How much evidence is enough? Do we need, for example, 60 minutes of video, or could we obtain sufficient validity and reliability with 15 minutes?
- What is the right evidence to support a specific claim? If, for example, we want to claim that a candidate can identify an instructional problem grounded in evidence, a video would not be appropriate. Instead, more relevant evidence is longitudinal student data that demonstrates trends over time. Examining this data allows evaluators to determine whether the candidate accurately analyzed multiple data points, identified a meaningful instructional need, and used evidence—not assumptions—to justify the focus of the improvement plan.

A task is something specific that we ask a candidate to do. Tasks generate evidence that supports the claims we wish to make about the candidate we are assessing. We have developed tasks specifically to generate evidence of components of practice, and each task measures one or more standards. Often, many points of evidence are required for each standard, and we have developed tasks to generate that evidence. Candidates are sometimes required to submit and discuss such things as student work, walk-through observation forms, assessments that measure students' progress, notes from meetings with colleagues, etc.

After we developed individual tasks, we reviewed them for consistency. It is important to make certain that the tasks, taken together, measure the various standards and that there is an alignment between the standards and assessment. Collectively, the tasks must be sufficient to provide an accurate picture of practice for the specific population of educators. The tasks must also provide the educators with the opportunity to demonstrate that practice. Finally, the tasks must be of equal rigor for the various candidates.

The ETS Performance Assessment for School Leaders (PASL)

This assessment is summative in nature consisting of tasks, commentary, and relevant artifacts. The tasks reflect national standards and state specific standards below. One of the tasks includes a 15-minute video submission. The assessment focuses on the knowledge and skills a candidate needs to become a principal.

- Professional Standards for Educational Leaders (PSEL) (PDF),
- National Educational Leadership Preparation (NELP) Program Recognition Standards: Building Level (PDF) and District Level (PDF).
- Georgia Educational Leadership Standards and Elements (PDF)
- Texas Principal Certificate Standards (PDF)

The development team first spent time examining the appropriate standards, determining which standards are most appropriate to cluster together, and then building an assessment around those clusters of standards. Development of the assessment also required a small-scale field test (tryout) as well as a large-scale field test (pilot) followed by a field test scoring session.

Candidates use an online submission system to create a portfolio in which they upload documents (task commentary and artifacts) as well as video.

The PASL assessment consists of three tasks. All tasks are completed during the clinical experience. ETS designed each task to address specific standards. However, the PASL assessment does not address every standard. Successful completion of coursework (as certified by the attending university faculty member) better addresses certain standards.

The Tasks

Each of the PASL assessment tasks includes a written commentary in which the candidate responds to a series of prompts and provides related artifacts. A development team of educators, with ETS facilitation, determined the prompts, the length of the written commentary, and the number of artifacts.

Task 1: Problem Solving in the Field

In this task, candidates demonstrate the ability to address and resolve a significant problem/challenge in the school that influences instructional practice and student learning. The task asks candidates to provide evidence in regard to their colleagues, the school and/or the community and to identify a problem/challenge that has implications for instructional practice and student learning.

Task 2: Supporting Continuous Professional Development

In Task 2, candidates demonstrate skills in establishing and supporting effective and continuous professional development with staff for the purpose of improved instruction and student learning. The candidate works with colleagues to develop a list of professional development needs and designs and implements a research-based professional development plan that addresses the most significant need(s).

Task 3: Creating a Collaborative Culture

In this task candidates demonstrate the ability to facilitate stakeholders' efforts to build a collaborative team within the school to improve instruction, student achievement and the school culture. A 15-minute video is required with this task.

Detailed requirements for each task may be found on the task requirements page of the performance assessments website.

Field Testing

For performance-based assessments we generally conduct two field tests. The first field test is to try out the assessment tasks with a smaller targeted group of individuals. The purpose is to determine at an early point in the development whether or not there are fundamental flaws in the desired approach of each assessment task. This allows for necessary corrective action before the second field test, which is larger and more formal.

In the description that follows, the initial field test is referred to as the "tryout" and second field test is referred to as the "pilot." We will use those terms in describing our field-testing routine throughout the text for these standards-based assessments.

Small-scale Field Test: The Tryout

We used the development team members to complete the initial tryout of the assessment tasks. In addition, we asked each team member to recruit at least two colleagues to also complete the tryout of the assessment tasks.

The purpose of the tryout was to obtain insight into the:

- clarity of the task directions
- value of the evidence elicited
- alignment of the evidence elicited to the directions and the standards
- thoroughness of the responses generated by the assessment directions
- alignment to the draft rubrics

Development Team Meeting: Tryout Review and Field Preparation

The primary focus at this meeting was to use the results of the tryout to finalize the assessment tasks and rubrics before the formal pilot of that assessment. Having this early opportunity to determine whether or not assessment directions are guiding candidates properly in building their response promotes a more efficient and effective pilot. The team also spent time on rubric development, cross-checking the tasks to the standards they are supposed to measure, and aligning the standards and tasks to the draft rubrics, as they were developed.

Large-scale Field Test: The Pilot

We convened a review panel to review the field-test plans, assessment directions, and draft rubrics prior to the pilot. Once the panel provided approval, we piloted the assessment using a larger number of participants than the tryout.

The pilot pool was selected to be diverse and representative across:

- gender
- race/ethnicity
- content area taught
- developmental level taught
- regional location

ETS delivered the pilot electronically, using our online submission system. The pilot allowed us to test the online submission system to answer the following questions specific to the standards-based assessments.

- Did we enter the tasks properly into the system? Do the directions and tasks display properly?
- Will users see the tasks in the proper sequence?
- Can candidates access and navigate the system easily?
- Are there any issues with entering evidence, submitting responses, or uploading attachments and artifacts?
- Are the directions and instructions clear and easy to follow for each section?
- Does the task measure skills and knowledge that are important for meeting standards? In other words, does the content reflect what candidates should know and be able to do relative to the standards?
- Are the tasks described equally accessible and applicable to all candidates within the certification areas regardless of ethnicity, gender, race, disability, or teaching context (e.g., urban, rural, and suburban; privileged; at-risk; homogeneous; diverse)?
- Do the activities associated with the task represent authentic (i.e., realistic) practices?
- Was the content and nature of the task fair?
- Was any information missing that should be added?
- Does each task adequately address the age range of the learners?
- Are the evaluation criteria equally applicable to all candidates within the certificate area regardless of ethnicity, gender, race, disability, or teaching context (e.g., urban, rural, and suburban; privileged; at-risk; homogeneous; diverse)?
- Are the described tasks free of language and/or content that reinforce stereotypes of educators or students (e.g., consider diversity issues related to gender, race, ethnicity, disability)?

In addition to submitting the pilot responses, the participants also supplied personal information by completing a Background Information Questionnaire. The participants also supplied information regarding ease of use of the system, clarity of task directions, time taken to complete the assessment, and additional information as determined with our clients.

The pilot responses underwent formative scoring, after which we made final revisions to the task directions and materials, as well as to the scoring materials, including rubrics.

After ETS scored the pilot responses as part of the formative scoring process, we gathered performance data for statistical analysis experts to analyze. These experts checked for inter-rater reliability, internal consistency reliability, and for issues related to difficulty (too hard, too easy).

Pilot Scoring and Formative Review

ETS recruited educators, including the development team members, to score the pilot responses and to validate the task directions as appropriate or make suggestions and modifications to the tasks and scoring rubrics in light of the responses. The emphasis during this process was not on the performance of the candidate, but rather on the performance of each task.

The pilot scoring session was conducted over multiple days. ETS staff spent several days training the raters and several days scoring the pilot responses in a round robin approach which allowed all raters to experience the three formative tasks. Raters double scored the various submissions.

Scoring took place electronically, using the Online Network for Evaluation (ONE), which is also used for operational scoring. A formative review based on the written feedback from the raters followed scoring.

Formative review refers to the analysis of the task requirements, guiding prompts and rubrics completed by the pilot scoring team after completion of scoring. Participants allow their experience in reviewing and scoring submissions to influence their judgments of the tasks. Once all participants complete the form, results are complied, and the development team and AD together review those comments and make appropriate changes to the directions, requirements, prompts, and rubrics.

The same form was sent to those students who submitted pilot responses. Again, the resulting information was compiled and shared with the pilot participants before completing the Formative Review Form. Both the development team and ETS Assessment Development staff reviewed these results.

At the end of pilot scoring, the team reviewed the data and information acquired and made final revisions to the tasks and rubrics to prepare them for the first official candidates.

Training Approach

During formative review, we tightly linked the development process and the scoring process. ETS Assessment Development staff have developed training protocol models that raters use in live scoring. Pairs of raters work together as a task-specific team that chooses benchmarks. The primary outcome of the formative review is to determine whether the task is working as designed. These activities attempt to answer questions like:

- Is there the anticipated connection between the standards, the evidence collected by the response, and what is valued in scoring the task?
- More specifically, formative review is the verification of the critical alignment between the standards of focus and the elicited evidence from educators through the assigned tasks. Do the tasks allow candidates to produce and document the necessary evidence that speaks to the standards in meaningful ways?
- Does the rubric both value the evidence and support raters, in making sense of the evidence that leads to consistent and valid judgments on the accomplishment of the teacher candidate relative to the specific task?

Evaluation of the Evidence

Assessment Development specialists generated and documented the professional judgments of review team. The Assessment Development specialists led these sessions and put the teams through a strict process considering feedback through a focused discussion of the explicit requirements of the ECD process. We asked team members to consider the evidence submitted by the pilot participants, whether tasks elicited the responses expected, and whether we maintained the intended connections between standards, task directions, and the scoring rubrics.

Standard Setting

To support the decision-making process for states in establishing a passing score (cut score) for the performance assessments, research staff from ETS designed and conducted a standard-setting study. Separate standard-setting studies were conducted for each assessment.

Each study provides a recommended passing score, which represents the combined judgments of a group of experienced educators. ETS provides a recommended passing score from the standard-setting study; state agencies are responsible for establishing the operational passing score in accordance with applicable regulations. ETS does not set passing scores; that is the state agency's responsibility.

Panel Formation

Standard-setting studies provide recommended passing scores, which represent the combined judgments of a group of experienced educators. For the standard-setting studies, state agencies recommended panelists with (a) experience as either educators in the particular area or college faculty who prepare educators in the area and (b) familiarity with the knowledge and skills required of beginning educators. ETS selects panelists to represent the diversity (race/ethnicity, gender, geographic setting, etc.) of the educator population in the state. Each panel includes 10-20 educators, the majority of whom are practicing, licensed educators in the area covered by the test.

Standard-Setting Process

Each study began with an in-depth discussion of the assessment, the tasks, scoring rubrics, and candidate handbooks. Panelists were asked to take notes on the tasks and steps, focusing on what is being measured and the challenge the tasks pose for beginning educators. Then panelists discussed the state standards, particularly the quality indicators. The quality indicates described a continuum of expected performance, from "candidate" to "distinguished educator." The candidate-level quality indicators define what is expected to be "just" qualified to begin practice. The just-qualified candidate description plays a central role in standard setting[3] (Perie, 2008); the goal of the standard-setting process is to identify the test score that aligns with this description[4] (Tannenbaum & Katz, 2013). The panelists' understanding of the assessments and the expectations for the just qualified candidate were crucial in the standard-setting process.

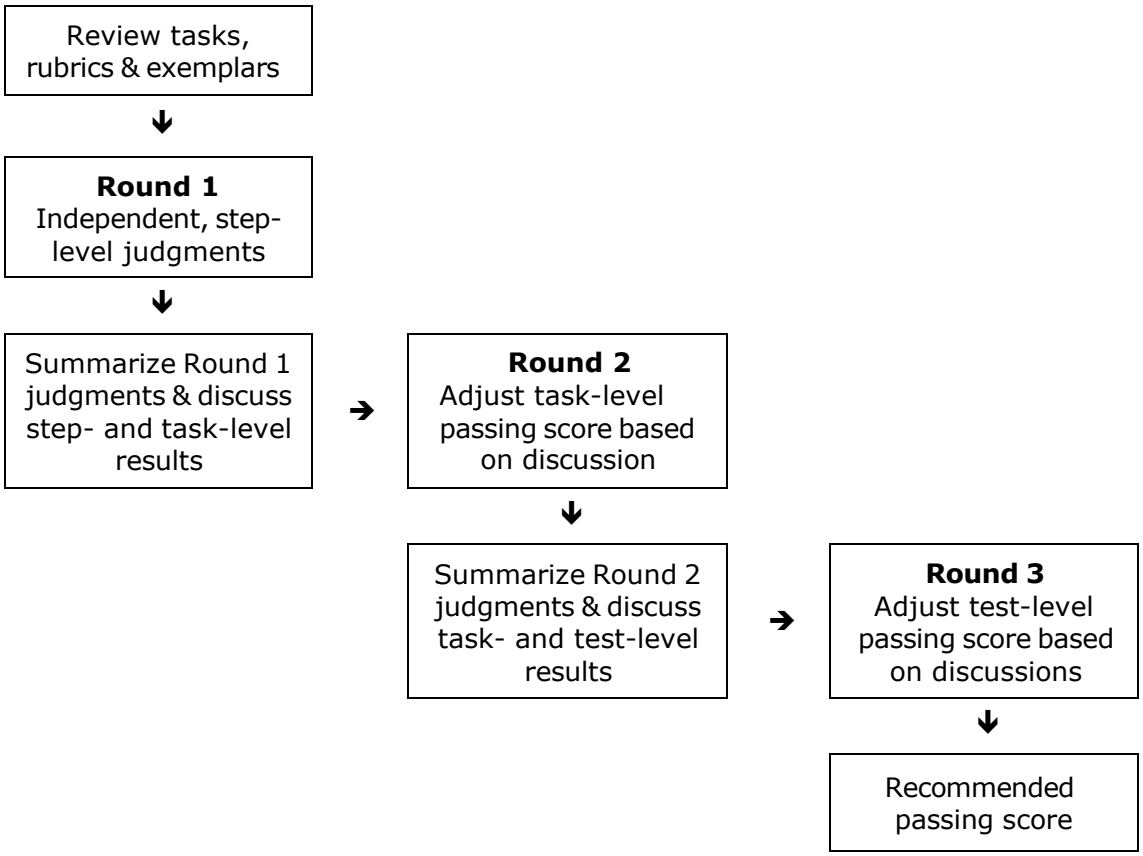
Standard-setting methods are selected based on the characteristics of the particular test. Given the structure of the performance assessments — each assessment consists of tasks and each task consists of steps — a standard-setting method was designed to reflect this nested structure. For the assessments, a variation on a multiple-round extended Angoff method[5] was used. Multiple rounds of judgments are collected, with panelist discussion between rounds.

The panelists made independent judgments at the step-level for their first round of judgments. They considered each step in a task, the rubrics, and exemplars. Then the panelists independently judge, for each step within the task, the score a just qualified candidate would likely receive. The task-level result from Round 1 was the simple sum of the recommended likely scores for each step.

For the second round of judgments, the panelists reviewed a summary of results from Round 1. They discuss their step-level judgments, then the task-level judgments. They are asked if the task-level score from Round 1 reflects the likely performance of the just qualified candidate, considering the various patterns of step scores that are likely to result in the task score. If sufficient field test data were available, the panel reviewed the distribution of candidate scores from the field test as a reasonableness check for the Round 1 judgments. The panelists then made a holistic task-level judgment as their Round 2 judgments. The assessment-level result from Round 2 is the weighted sum of the recommended likely scores for each task.

For the third round of judgments, the panelists reviewed a summary of results from Round 2. They discuss their task-level judgments, then the assessment-level judgments. They are asked if the assessment-level score from Round 2 reflects the likely performance of the just qualified candidate, considering the various patterns of task scores that are likely to result in an assessment score. Again, if sufficient field test data were available, the panel reviewed the distribution of candidate scores from the field test as a reasonableness check for the Round 2 judgments. The panelists then made a holistic assessment-level judgment as their Round 3 judgments. The results from Round 3 were the final judgments and the panel's average was the recommended passing score.

Figure 2. Standard-Setting Process



Standard-Setting Reports

Following the study, ETS created a technical report for each assessment. The technical report describes the content and format of the performance assessment, the standard-setting processes and methods, and the results of the standard-setting study. The standard-setting technical reports can be made available upon request.

Scoring Methodology

All PASL assessment tasks are centrally scored by trained educators using ETS's Online Network for Evaluation (ONE). Within a task, each step receives a score. Scores for the steps are combined to determine the task score.

Scores for the three tasks are then combined to produce an overall score. The final task is weighted so that it contributes more to the overall score than the other two tasks. The passing or cut scores were set by standard-setting panels composed of state educators (see the section on Standard Setting above). There are no passing scores for the individual tasks.

Centralized Scoring Model

ETS employs and trains the raters who score the performance assessments using ETS's Online Network for Evaluation (ONE). Raters who are faculty at EPPs are not permitted to score candidates attending their EPP, and mentor teachers do not score submissions from candidates completing their clinical experience in their districts. Submissions are scored using a targeted double scoring model to monitor scoring quality. Each task is single scored, but if a total task score falls within a pre-defined critical range, at least one additional rater will independently score the task (Miao et al., 2023). ETS provides full support to the scoring process — scoring materials (e.g., rubrics), training materials, and scoring oversight (e.g., scoring leaders).

Scoring Process

When scoring tasks, raters review the written commentary the candidate entered in the online submission system and the artifacts that the candidate uploaded, including the video.

Calculating Step Scores

For the PASL assessment there are 12 steps. Step scores are determined using a four-point rubric. Score levels for each rubric are defined as follows:

Score	Quantitative and Qualitative Elements of Evidence
Score of 4	Consistent and thorough
Score of 3	Effective and appropriate
Score of 2	Partial and inconsistent
Score of 1	Minimal and ineffective
0	Step's textboxes are blank; artifacts missing for a step or inappropriate; response is off-topic or is plagiarized

The rubric documents contain the task-specific rubrics used during scoring to evaluate the elements of the evidence provided for each step.

PASL Assessment

- Task 1 Rubric (PDF)
- Task 2 Rubric (PDF)
- Task 3 Rubric (PDF)

Steps that are determined to be nonscorable receive a score of zero.

Calculating Task Scores

Step scores are summed to determine the task score for each of the three tasks. The score for the video task (Task 3 for the PASL assessment) is multiplied by two to reflect the double weighting of the task. Tasks that are not submitted receive a score of zero.

At least three raters contribute to scoring the assessment.

Calculating the Overall Assessment Score

The three task scores are summed to determine the overall assessment score. As noted above, the score for the video task is doubled.

Score Reports

As candidates submit each of their completed tasks into the online submission system, raters score the tasks within a relatively short turnaround time. Score reports are available via the candidate's online account approximately three or four weeks after task deadlines. The candidate and the EPP will have access to the scores on each task and can use them to engage in reflective practice. This will help candidates become more skillful practitioners as they continue through the clinical experience. The final score report, including a composite score of all three tasks, is available to candidates through their online account approximately three to four weeks after the deadline date for the final task. EPPs and the state agencies access candidate score reports through the ETS Data Manager (EDM). If a candidate is not successful, the submission schedule will afford the teacher candidate the opportunity to resubmit any task(s).

Quality Assurance Measures

All raters are carefully trained and follow strict scoring procedures. If a candidate's cumulative score falls below the designated passing score for the assessment, they are eligible to resubmit any or all three tasks for a fee during the resubmission window immediately following their original submission.

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 our tests to understand what are considered their proper uses. The booklet *Proper Use of The ETS Professional Educator Programs Assessments (PDF)* defines proper test use as adequate evidence to support the intended use of the test and to support the decisions and outcomes rendered on the basis of test scores.

Proper assessment use is a joint responsibility of ETS as the test developer, and of states, agencies, associations, and institutions of higher education as the test users. The *Praxis* program is responsible for developing valid and fair assessments in accordance with technical guidelines established by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014).

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

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

Psychometric Properties

Introduction

ETS Psychometric Analysis and Research (PAR) division has established procedures to support the development of valid and reliable interpretation of test scores for the PASL Assessment. Software developed at ETS provides rigorously tested routines to produce task level and test level statistics at various stages of each test administration, which enables candidates to receive feedback and make improvements to their responses before final assessment scores are reported.

After the initial submission, task score distributions are examined to ensure scores are within range and reasonable. Task ratings from different raters are compared and rater agreement statistics are calculated to monitor the rating quality and to provide feedback for training and/or improvement of the scoring process. Total test score distributions are also examined to ensure scores are within range and reasonable, and to identify any anomaly for investigation. Test reliability is estimated to evaluate the overall quality of the assessment. Based on the task level feedback from the initial submission, candidates can revise their work and resubmit any or all tasks. The task and test level analyses are repeated for resubmissions before the release of the final task and test scores.

Test Statistics

Reliability and the Standard Error of Measurement

Reliability is a measure of the extent to which scores are expected to be consistent over time. Reliability coefficients may range from 0 to 1. The higher the reliability coefficient, the more likely individuals would be to obtain very similar scores if they were retested. In this report Cronbach's alpha (Cronbach, 1951) is used as a measure of internal consistency of the assessments. The Cronbach's alpha of the total assessment score is calculated on the basis of the task scores (i.e., with a test length of 3).

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 or at administrations of the same test without benefiting from practice or being hampered by fatigue. The SEM is computed from the alpha reliability estimate (r_x) and the standard deviation (SD_x) of the assessment scores; the formula is included in Appendix A.

Inter-Rater Reliability and the Standard Error of Scoring

The inter-rater reliability coefficient describes the reliability of the scoring process when constructed response items are scored independently by two raters. It is an estimate of the correlation between the scores resulting from two independent replications[6] of the same scoring process.

The standard error of scoring (SES) is an estimate of the standard deviation of the distribution of observed scores around a theoretical true score. The SES can be interpreted as an index of expected variation if the same test taker's responses were scored repeatedly by different raters on the same test responses. The SES is computed from the inter-rater reliability estimate ($r_{scoring}$) and the standard deviation (SD_x). The multi-step calculation process is included in Appendix A.

Reliability of Classification: Decision Accuracy and Decision Consistency

When a cut score is used for pass or fail decisions, reliability of classification measures also need to be evaluated. We use the computer program RELCLASS, which operationalizes a statistical method developed by Livingston and Lewis (1995). This method uses information from the administration of one test form (i.e., distribution of scores, the minimum and maximum possible scores, the cut points used for classification, and the alpha reliability coefficient) to estimate two kinds of statistics, "decision accuracy" and "decision consistency".

Decision accuracy refers to the extent to which the classifications of test takers based on their scores on the test form agree with the classifications made on the basis of the classifications that would be made if the test scores were perfectly reliable. Decision consistency refers to the agreement between these classifications based on two non-overlapping, equally difficult forms of the test.

Decision consistency values are always lower than the corresponding decision accuracy values, because in decision consistency, both of the classifications are based on scores that depend on which form of the test the candidate took. In decision accuracy, only one of the classifications is based on a score that can vary depending on the test form. It is not possible to know which candidates were accurately classified, but it is possible to estimate the proportion of the candidates who were accurately classified. Similarly, it is not possible to know which candidates would be consistently classified if they were retested during another administration, but it is possible to estimate the proportion of the candidates who would be consistently classified.

Score Reporting

Score reporting is the process by which tests are scored and test results are reported to candidates, institutions, and state agencies.

Candidate Score Reports – PASL Assessment

Candidates access their score reports via their online accounts. The PASL assessment score report contains valuable information, including:

- a summary page indicating the score for each task and the cumulative score for the assessment
- a detailed page for each task indicating scores for each step within a task and feedback for each step score

See a PASL Assessment Score Report (PDF).

Pass/Fail

Scores for each task are summed to determine the overall assessment score. The passing score is established by the state agency.

Reviewing the Feedback for Step Scores

Each task of the assessment contains four steps.

The candidate score report includes feedback for each step score within a task. This feedback:

- guides test takers to improve the quality of evidence in their step responses
- addresses the possible qualitative and quantitative level of the evidence provided in the step responses
- is connected to the language of the task rubrics and the language of the guiding prompts
- is helpful in deciding whether or not to resubmit a task

When making decisions about resubmission, candidates are encouraged to:

- read, review and reflect on the feedback provided with each step score
- reread the rubric at the score level obtained as well as at the next higher level(s)
- view the feedback for all steps and score levels of the PASL assessment
    - Task 1 Score Report Feedback (PDF)
    - Task 2 Score Report Feedback (PDF)
    - Task 3 Score Report Feedback (PDF)
- read some appropriate examples from the Library of Examples

Steps with a Score of Zero

If a score report contains a step score of zero, the step response may:

- be blank.
- not address any of the guiding prompts for the step.
- be attached as a standalone document rather than directly in the textbox provided.
- have a technical difficulty with the artifact attachment (e.g., the artifact is corrupt or will not open, is unreadable and/or indecipherable, or contains only hyperlinks).
- have a video artifact that was edited (e.g., eliminating unwanted sections within segments, adding footage, adding audio-recorded material from another device, fade-ins, and/or fade-outs), resulting in every step receiving a 0
- have none of the required artifacts acceptable or attached to any of the Task textboxes.

Score Reports for EPPs and the State Departments

EPPs and the state agency access candidate score reports through the ETS Data Manager (EDM).

Information provided through the EDM includes candidate demographic information; Current Scores including Task and Step Scores, the cumulative score for the assessment, and Highest Scores earned to date. The state agency also has access to aggregate score data by EPP.

References

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Educational Testing Service (2004). *Questions to Ask About Teacher Testing*. Princeton, NJ.

Educational Testing Service (2014). *ETS Standards for Quality and Fairness*. Princeton. NJ.

Educational Testing Service (2020). *Proper Use of The PRAXIS Series and Related Assessments*. Princeton, NJ.

Educational Testing Service (2022). *ETS Guidelines for Fairness Review of Assessments*. Princeton, NJ.

Educational Testing Service (September 2023). *Technical Manual for the Praxis® Tests and Related Assessments*. https://www.ets.org/pdfs/praxis/technical-manual.pdf

Joint Committee on Testing Practices (2004). *Code of Fair Testing Practices in Education*. Washington, DC.

Knapp, J., and 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.

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

Miao, J., Sinharay, S., Kelbaugh, C., Cao, Y., & Wang, W. (2023). *Evaluating Targeted Double Scoring for the Performance Assessment for School Leaders Using Imputation and Decision Theory*. ETS Research Report Series. (ETS RR-23-01). https://onlinelibrary.wiley.com/doi/10.1002/ets2.12363

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*, 15–29.

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

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

Shrout, P. E. and Fleiss, J. L. (1979) Intraclass correlations: use in assessing rater reliability. *Psychological Bulletin*, Vol. 86, No. 2, 420-428.

Tannenbaum, R. J., and 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.

Tannenbaum, R. J. & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology*. Washington, District of Columbia: American Psychological Association.

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.

Appendix: Statistical Characteristics of the PASL Assessment

Table 1 in this section provides important scoring and statistical information for the PASL assessment. Notes at the end of the table provide more information about the data included. Table 3 provides summary statistics for subgroups by gender and ethnicity.

Table 1: Statistical Characteristics for the PASL Assessment

2024-2025	PASL
Sample Size	2,567
Possible Range	0-64
Observed Range	0-55
Median	46
Mean	42.35
Standard Deviation	11.77
Reliability: Alpha	0.87
Standard Error of Measurement	4.27
Reliability: Inter-Rater	0.81
Standard Error of Scoring	2.22
Cut Score	42
Decision Accuracy	0.86
Decision Consistency	0.84

- **Sample Size** — The number of people taking the test and with a valid reported score.
- **Possible Range** — The lowest to the highest raw score possible on any edition of the test.
- **Observed Range** — The actual maximum and minimum observed raw scores for a given form of a test.
- **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.
- **Mean** — The arithmetic average, calculated for the scores obtained by the group of test takers.
- **Standard Deviation** — The amount of variability among the scores obtained by the group of test takers.
- **Alpha Reliability** — Cronbach's alpha is calculated as a measure of the consistency of scores. It focuses on the extent to which differences in test scores reflect true differences in the knowledge, ability, or skills being tested. Reliability coefficients range from 0 to 1 and is calculated using the following formula:

$$\hat{\alpha} = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{s_C^2}\right] \quad (1)$$

where,

k is the number of criteria, which is 3 for PASL,

s_i^2 is the variance of scores for Task *i*, and

s_C^2 is the variance of total assessment score.

- **Standard Error of Measurement** (SEM) — The standard error of measurement (SEM) is a test statistic to characterize the reliability of the scores of a group of test takers. A test taker's score on a single administration of a test will differ somewhat from the score the test taker would receive on another occasion. The more consistent a test taker's scores are from one testing to another, the smaller the SEM. The SEM is calculated using the formula below:

$$SEM = SD_{total\ assessment\ score} * \sqrt{1-\alpha} \quad (2)$$

- **Inter-rater Reliability and Standard Error of Scoring** (SES) — The inter-rater reliability coefficient describes the reliability of the scoring process when constructed response items are scored independently by two raters. The standard error of scoring (SES) is an estimate of the standard deviation of the distribution of observed scores around a theoretical true score. The calculation uses a multi-step process described below. Since PASL follows a targeted double scoring procedure (see Scoring Procedures section), the calculation is based on a subset of test candidates who have at least one task scored by two raters.

  Step 1: Calculate Intra-class Correlation for Each Task
  The intra-class correlations (ICC) are calculated for each task. Since the first two ratings are used for the calculation, the ICC values would be somewhat underestimated. The intra-class correlation is a measure of agreement among ratings (Shrout & Fleiss, 1979). The ICC ranges from 0 to 1, with 1 indicating the two ratings agree perfectly for all test takers and 0 indicating no relationship. The intra-class correlation is derived from the Analysis of Variance (ANOVA) framework. Table 2 shows the decomposition of the variance for each criterion when there are two ratings for each response. In this table, the subject represents the candidate's response and *n* is the number of test takers.

Table 2. Analysis of Variance Framework

Source of Variation	df	Mean Square
Between subject	n-1	BMS
Within subject	n	WMS

The intra-class correlation is calculated as:

$$r_{icc} = \frac{BMS - WMS}{BMS + WMS}, \quad (3)$$

where BMS is the Between Subject Mean Square, which is calculated as:

$$BMS = \frac{2*\sum_{i=1}^{n}(\bar{X}_{i.} - \bar{X}_{..})^2}{n-1} \quad (4)$$

and WMS is the Within Subject Mean Square, which is calculated as:

$$WMS = \frac{\sum_{i=1}^{n}(X_{i1}^2 + X_{i2}^2 - 2\bar{X}_{i.}^2)}{n} \quad (5)$$

In the formula above, X_{i1} is the first rating for the response of test taker *i*, X_{i2} is the second rating for test taker *i*, $\bar{X}_{i.}$ is the average rating for test taker *i*, and $\bar{X}_{..}$ is the overall average of the ratings across all test taker for all *n* test taker:

$$\bar{X}_{i.} = \frac{X_{i1} + X_{i2}}{2} \quad (6)$$

$$\bar{X}_{..} = \frac{\sum_{i=1}^{n}(X_{i1} + X_{i2})}{2n} \quad (7)$$

Step 2: Calculate Reliability of Combined Item Score (CIS) for Each Task
A combined item score (CIS) is the average of two ratings assigned by different raters to the same response.

$$r_{CIS} = \frac{2r_{icc}}{1 + r_{icc}} \quad (8)$$

Step 3: Calculate the Variance of Error of Scoring for Each Task

$$\sigma_{e_{CIS}}^2 = \sigma_{x_{CIS}}^2 (1 - r_{CIS}) \quad (9)$$

$\sigma_{x_{CIS}}^2$ is the variance of the combined item score (CIS), i.e., the squared SD of the CIS.

Step 4: Calculate the Variance of Error of Scoring for the Entire Test

$$\sigma_{e_{scoring}}^2 = \sum w^2 \sigma_{e_{CIS}}^2 \quad (10)$$

This is the weighted sum of the three task variances from step 3. *W* is the weight for each task score in computing the total test score, i.e., 1, 1 and 2 for Task 2, Task 3 and Task 4 respectively.

Step 5: Calculate the Standard Error of Scoring for the Entire Test

$$SES = \sqrt{\sigma^2_{\hat{e}_{scoring}}} \qquad (11)$$

Step 6: Calculate the Reliability of Scoring

$$r_{scoring} = 1 - \frac{\sigma^2_{\hat{e}_{scoring}}}{\sigma^2_{total\ assessment\ score}} \qquad (12)$$

Table 3: Summary of PASL Assessment Scores by Gender and Ethnicity

2024-2025	N	Mean	Std	Min	Max	Median
Total	2567	42.35	11.77	0	55	46
No Response	9	41.22	11.57	9	48	44
Female	1959	43.09	10.64	0	55	46
Male	598	39.95	14.65	0	54	45
Prefer not to answer	1	43	0	43	43	43
Unspecified Ethnicity	116	43.68	10	0	52	46
African American	531	41.08	12.55	0	53	45
Asian or Pacific Islander	29	44.24	8.88	0	50	47
Hispanic	554	42.85	11.39	0	54	46
American Indian or Alaskan Native	9	36.67	15.73	0	48	44
White	1237	42.51	11.72	0	55	46
Other	20	42.95	10.57	0	48	46
Two or More Races	71	42.68	12.09	0	52	46