Multistate Standard-Setting Technical Report for the

Praxis® Russian: World Language (5671)

Student and Teacher Assessments: Validity and Test Use

ETS

Princeton, New Jersey

April 2022

Executive Summary

To support the decision-making process of education agencies establishing a passing score (cut score) for the *Praxis*® Russian: World Language (5671) test, research staff from Educational Testing Service (ETS) designed and conducted a multistate standard-setting study (Tannenbaum, 2011, 2012).

Participating States

Panelists from four states were recommended by their respective education agencies. The education agencies recommended panelists with (a) experience as either Russian language teachers or college faculty who prepare those Russian language teachers and (b) familiarity with the knowledge and skills required of beginning Russian language teachers.

Recommended Passing Score

ETS provides a recommended passing score from the multistate standard-setting study to help education agencies determine an appropriate operational passing score. For the *Praxis* Russian: World Language test, the recommended passing score is 39 out of a possible 98 raw-score points. The scale score associated with a raw score of 39 is 130 on a 100–200 scale.

Introduction

To support the decision-making process for education agencies establishing a passing score (cut score) for the *Praxis*® Russian: World Language (5671) test, research staff from ETS designed and conducted a multistate standard-setting study (Tannenbaum, 2011, 2012) in March 2022. Education agencies[1] recommended panelists with (a) experience as either Russian language teachers and (b) familiarity with the knowledge and skills required of beginning Russian language teachers. Four states (Table 1) were represented by seven panelists. (See Appendix A for the names and affiliations of the panelists.)

Table 1
Participating States and the Number of Panelists

Arkansas (1 panelist)	Maryland (4 panelists)
Kansas (1 panelist)	Utah (1 panelist)

The following technical report contains three sections. The first section describes the content and format of the test. The second section describes the standard-setting processes and methods. The third section presents the results of the standard-setting study.

ETS provides a recommended passing score from the multistate standard-setting study to education agencies. In each state, the department of education, the board of education, or a designated educator licensure board is responsible for establishing the operational passing score in accordance with applicable regulations. This study provides a recommended passing score, which represents the combined judgments of a group of experienced educators. Each state may want to consider the recommended passing score but also other sources of information when setting the final *Praxis* Russian: World Language passing score (see Geisinger & McCormick, 2010). A state may accept the recommended passing score, adjust the score upward to reflect more stringent expectations, or adjust the score downward to reflect more lenient expectations. There is no *correct* decision; the appropriateness of any adjustment may only be evaluated in terms of its meeting the state's needs.

[1] States and jurisdictions that currently use *Praxis* tests were invited to participate in the multistate standard-setting study.

Two sources of information to consider when setting the passing score are the standard error of measurement (SEM) and the standard error of judgment (SEJ). The former addresses the reliability of the *Praxis* Russian: World Language test score and the latter, the reliability of panelists' passing-score recommendation. The SEM allows states to recognize that any test score on any standardized test—including a *Praxis* Russian: World Language test score—is not perfectly reliable. A test score only *approximates* what a candidate truly knows or truly can do on the test. The SEM, therefore, addresses the question: How close of an approximation is the test score to the *true* score? The SEJ allows states to gauge the likelihood that the recommended passing score from the current panel would be similar to the passing scores recommended by other panels of experts similar in composition and experience. The smaller the SEJ, the more likely that another panel would recommend a passing score consistent with the recommended passing score. The larger the SEJ, the less likely the recommended passing score would be reproduced by another panel.

In addition to measurement error metrics (e.g., SEM, SEJ), each state should consider the likelihood of classification errors. That is, when adjusting a passing score, policymakers should consider whether it is more important to minimize a false-positive decision or to minimize a false-negative decision. A false-positive decision occurs when a candidate's test score suggests that they should receive a license/certificate, but their actual level of knowledge/skills indicates otherwise (i.e., the candidate does not possess the required knowledge/skills). A false-negative decision occurs when a candidate's test score suggests that they should not receive a license/certificate, but they actually do possess the required knowledge/skills. States needs to consider which decision error is more important to minimize.

Overview of the Praxis® Russian: World Language Test

The *Praxis*® Russian: World Language *Study Companion* document (ETS, in press) describes the purpose and structure of the test. In brief, the test measures whether entry-level Russian language teachers have the knowledge/skills believed necessary for competent professional practice.

The 3-hour assessment measures contains 75 selected-response items[2,3] and 6 constructed-response items[4] covering five content areas*: Interpretive Listening, including embedded linguistic content (approximately 30 items), Interpretive Reading, including embedded linguistic content* (approximately 30 items), *Cultural Knowledge* (approximately 15 items), *Interpersonal and Presentational Writing* (approximately 3 items), and *Interpersonal and Presentational Speaking* (approximately 3 items).[5] The reporting scale for the *Praxis* Russian: World Language test ranges from 100 to 200 scale-score points.

Processes and Methods

The design of the standard-setting study included an expert panel. Before the study, panelists received an email explaining the purpose of the standard-setting study and requesting that they review the content specifications for the test. This review helped familiarize the panelists with the general structure and content of the test.

The standard-setting study began with a welcome and introduction by the meeting facilitator. The facilitator described the test, provided an overview of standard setting, and presented the agenda for the study. Appendix B shows the standard-setting study agenda.

Reviewing the Test

The standard-setting panelists first took the test and then discussed the content measured. This discussion helped bring the panelists to a shared understanding of what the test does and does not cover, which serves to reduce potential judgment errors later in the standard-setting process.

The test discussion covered the major content areas being addressed by the test. Panelists were asked to remark on any content areas that would be particularly challenging for entry-level Russian language teachers or areas that address content particularly important for entry-level Russian language teachers.

[2] Thirteen of the 75 selected-response items are pretest items and do not contribute to a candidate's score.
[3] Six, non-scored, selected-response items are included as a Listening Practice section for candidates. They were not included in the standard setting operational judgments.
[4] One, non-scored, constructed-response item is included as a Writing Practice section for candidates. It was not included in the standard setting operational judgments.
[5] The number of items for each content area may vary slightly from form to form of the test.

Defining the Just-Qualified Candidate

Following the review of the test, panelists described the just-qualified candidate. The *just-qualified candidate description* plays a central role in standard setting (Perie, 2008); the goal of the standard-setting process is to identify the test score that aligns with this description.

The panel created a description of the just-qualified candidate—the knowledge/skills that differentiate a *just-qualified* from a *not quite-qualified* candidate. To create this description, the panel first split into smaller groups to consider the just-qualified candidate. Then they reconvened and, through whole-group discussion, determined the description of the just-qualified candidate to use for the remainder of the study.

The written description of the just-qualified candidate summarized the panel discussion in a list format. The description was not intended to describe all the knowledge and skills of the just-qualified candidate but only highlight those that differentiate a *just-qualified candidate* from a *not-quite-qualified* candidate. The written description was distributed to panelists to use during later phases of the study (see Appendix C for the just-qualified candidate description).

Panelists' Judgments

The *Praxis* Russian: World Language test includes both dichotomously-scored (i.e., selected-response items) and constructed-response items. Panelists received training in two distinct standard-setting approaches: one standard-setting approach for the dichotomously-scored items and another approach for the constructed-response items.

A panel's passing score recommendation is the mean of the interim passing scores recommended by each of the panelists for (a) the dichotomously-scored items and (b) the constructed-response items. As with scoring and reporting, the panelists' judgments for the constructed-response items were weighted such that they contributed 36% of the overall score.

Dichotomously-scored items. The standard-setting process for the dichotomously-scored items was a probability-based Modified Angoff method (Brandon, 2004; Hambleton & Pitoniak, 2006). Using this method, each panelist judged each item on the likelihood (probability or chance) that the just-qualified candidate would answer the item correctly. Panelists made their judgments using the following rating scale: 0, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95, 1. The lower the value, the less likely it is that the just-qualified candidate would answer the item correctly because the item is difficult for the

just-qualified candidate. The higher the value, the more likely it is that the just-qualified candidate would answer the item correctly.

Panelists were asked to approach the judgment process in two stages. First, they reviewed both the description of the just-qualified candidate and the item and determined the probability that the just-qualified candidate would answer the question correctly. The facilitator encouraged the panelists to consider the following rules of thumb to guide their decision:

- Items in the 0 to .30 range were those the just-qualified candidate would have a *low chance* of answering correctly.
- Items in the .40 to .60 range were those the just-qualified candidate would have a *moderate chance* of answering correctly.
- Items in the .70 to 1 range were those that the just-qualified candidate would have a *high chance* of answering correctly.

Next, panelists decided how to refine their judgment within the range. For example, if a panelist thought that there was a *high chance* that the just-qualified candidate would answer the question correctly, the initial decision would be in the .70 to 1 range. The second decision for the panelist was to judge if the likelihood of answering it correctly is .70, .80, .90, .95 or 1.

After the training, panelists made practice judgments and discussed those judgments and their rationales. All panelists completed a post-training evaluation to confirm that they had received adequate training in the Modified Angoff method and felt prepared to continue; the standard-setting process continued only if all panelists confirmed their readiness.

Constructed-response items. An Extended Angoff method (Cizek & Bunch, 2007; Hambleton & Plake, 1995) was used for the constructed-response items. For this portion of the study, a panelist decided on the assigned score value that would most likely be earned by the just-qualified candidate for each constructed-response item. Panelists were asked first to review the definition of the just-qualified candidate and then to review the constructed-response item and its rubric. The rubric for a constructed-response item defines (holistically) the quality of the evidence that would merit a response earning a particular score. During this review, each panelist independently considered the level of knowledge/skill required to respond to the constructed-response item and the features of a response that would earn a particular score, as defined by the rubric. Each panelist decided on the score most likely to be earned by the just-qualified candidate from the possible values a test taker can earn.

A test-taker's response to a constructed-response item is independently scored by two raters, and the sum of the raters' scores is the assigned score[6]. Therefore, possible scores, range from zero (i.e., both raters assigned a score of 0) to six (i.e., both raters assigned a score of 6). For their ratings, each panelist decided on the score most likely to be earned by a just-qualified candidate from the following possible values: 0, 1, 2, 3, 4, 5, or 6. For each of the constructed-response items, panelists recorded the score (from 0 to 6) that a just-qualified candidate would most likely earn.

After the training, panelists made practice judgments and discussed those judgments and their rationales. All panelists completed a post-training evaluation to confirm that they had received adequate training in the Extended Angoff method and felt prepared to continue; the standard-setting process continued only if all panelists confirmed their readiness.

Multiple Rounds. Following this first round of judgments (*Round 1*), item-level feedback was provided to the panel. The panelists' judgments were displayed for each item and summarized across panelists. Item-level data were highlighted to show when panelists converged in their judgments or diverged in their judgments. For the dichotomously-score items, this meant that at least two-thirds of the panelists' judgments were in the same difficulty range. For the constructed-response items, this meant that at least two-thirds of the panelists' judgments indicated the same score most likely earned by a just-qualified candidate.

The panelists discussed their item-level judgments. These discussions helped panelists maintain a shared understanding of the knowledge/skills of the just-qualified candidate and helped to clarify aspects of items that might not have been clear to all panelists during the Round 1 judgments. The purpose of the discussion was not to encourage panelists to conform to another's judgment, but to understand the different relevant perspectives among the panelists.

In Round 2, panelists discussed their Round 1 judgments and were encouraged by the facilitator (a) to share the rationales for their judgments and (b) to consider their judgments in light of the rationales provided by the other panelists. Panelists recorded their Round 2 judgments only for items when they wished to change a Round 1 judgment. Panelists' final judgments for the study, therefore, consist of their Round 1 judgments and any adjusted judgments made during Round 2.

[6] If the two raters' scores differ by more than one point (non-adjacent), the Chief Reader for that item assigns the score, which is then doubled.

Results

Expert Panels

Table 2 presents a summary of the panelists' demographic information. The panel included seven educators representing four states. (See Appendix A for a listing of panelists.) Three panelists were teachers, one was a school testing coordinator, two were college faculty, and one was a college administrator (director and acting director of two Masters programs). The school testing coordinator and one of the teachers indicated that they are also supervising or mentoring other Russian language teachers.

Table 2
Panel Member Demographics

Background Survey Question	Number	Percent
What is your current position?	**N**	**%**
Teacher	3	38
School Testing Coordinator	1	14
College Faculty	2	29
MAT Director and MEd Acting Director	1	14
How do you describe yourself (i.e., race/ethnicity)?	**N**	**%**
White	7	100
What is your gender?	**N**	**%**
Female	7	100
Male	0	0
Are you currently certified to teach the Russian language in your state?*	**N**	**%**
Yes	2	50
No	2	50
Are you currently teaching the Russian language in your state?*	**N**	**%**
Yes	4	100
No	0	0
Are you currently supervising or mentoring other teachers of the Russian language?*	**N**	**%**
Yes	2	50
No	2	50
At what P–12 grade level are you currently teaching the Russian language?*	**N**	**%**
Elementary (P - 5 or P - 6)	3	75
Middle School (6 - 8 or 7 - 9)	1	25

(table continues on the next page)

Table 2 (continued from the previous page)
Panel Member Demographics

Including this year, how many years of experience do you have teaching the Russian language?*	**N**	**%**
3 years or less	0	0
4–7 years	4	100
8–11 years	0	0
12–15 years	0	0
16 years or more	0	0
Which best describes the location of your P–12 school?*	**N**	**%**
Urban	4	100
Suburban	0	0
Rural	0	0
Not currently working at the P–12 level	0	0
If you are college faculty, are you currently involved in the training/ preparation of Russian language teachers?	**N**	**%**
Yes	0	0
No	2	29
Not college faculty	5	71

Note: Questions indicated with an asterisk (*) were not presented to college faculty, administrators, or department heads.

Standard-Setting Judgments

Table 3 summarizes the standard-setting judgments of each panelist and shows the passing score recommendations of each panelist at each round—the number of raw points needed to "pass" the test. The recommendations are the raw score points needed out of a maximum of 98.

Table 3
Raw Score Recommendation of Each Panelist by Round of Judgments

Panelist	Round 1	Round 2
1	32.85	36.25
2	47.25	36.70
3	50.45	41.10
4	41.40	37.20
5	74.85	64.90
6	5.50	7.50
7	47.65	45.15

Table 4 shows the summary statistics at each round of judgment. The mean represents the panel's passing score recommendation at each round. Table 4 also includes the standard deviation and the standard error of judgment (SEJ). The SEJ is one way of estimating the reliability or consistency of a

panel's standard-setting judgments. It indicates how likely it would be for several other panels of educators similar in makeup, experience, and standard-setting training to the current panel to recommend the same passing score on the same form of the test. (Appendix D provides the technical notes, which further describe the SEJ.)

Table 4

Summary Statistics by Round of Judgments

Statistic	Round 1	Round 2
Mean	42.85	38.40
Minimum	5.50	7.50
Maximum	74.85	64.90
SD	20.90	16.93
SEJ	7.90	6.40

Data from Panelists 5 and 6 were detected to be outliers (High, 2000; see Appendix D). However, ETS does not recommend that their data be removed from the panel recommendation. Based on a report from the panel facilitator, the panelists were believed to be following the standard-setting process faithfully. Throughout the standard-setting, panelists are encouraged to consider the perspectives of their colleagues but that were not required to agree with their judgments.

Round 1 judgments are made without discussion among the panelists. The most variability in judgments, therefore, is typically present in the first round. Round 2 judgments, however, are informed by panel discussion; thus, it is common to see a decrease both in the standard deviation and SEJ. This decrease—indicating convergence among the panelists' judgments—was observed (see Table 4).

The Round 2 mean score is the panel's final recommended passing score. The panel's passing score recommendation for the *Praxis* Russian: World Language test is 38.40 (out of a possible 98 raw-score points). The value was rounded to the next highest whole number, 39, to determine the functional recommended passing score. The scale score associated with 39 raw points is 130.

The conditional standard error of measurement (CSEM) around the recommended passing score is 4.71 raw points. A standard error represents the uncertainty associated with a test score (See Appendix D for further information about the CSEM.) Table 5 shows the raw scores and the scale scores associated with one and two CSEM below and above the recommended passing score.

Table 5

Scores 1 and 2 CSEM Around the Recommended Passing Score (RPS)

Scores	Raw Score Points out of 98	Praxis Scale Score Equivalent
RPS - 2 CSEM	30	119
RPS - 1 CSEM	35	125
RPS	**39**	**130**
RPS +1 CSEM	44	137
RPS +2 CSEM	49	143

Notes. CSEM = conditional standard error(s) of measurement. The CSEM of the recommended passing score is 4.71 raw points. The unrounded CSEM value is added to, or subtracted from, the rounded passing-score recommendation. The resulting values are rounded up to the next-highest whole number and then converted to scale scores.

Final Evaluations

The panelists completed an evaluation at the conclusion of the standard-setting study. The evaluation asked the panelists to provide feedback about the quality of the standard-setting implementation and the factors that influenced their decisions. The responses to the evaluation provided evidence of the validity of the standard-setting process, and, as a result, evidence of the reasonableness of the recommended passing score.

Panelists were shown the panel's recommended passing score after Round 2 and asked, in the evaluation, (a) how comfortable they are with the recommended passing score and (b) if they think the score was *too high*, *too low*, or *about right*. A summary of the final evaluation results is presented in Appendix E.

All panelists *strongly agreed* that they understood the purpose of the study and that the facilitator's instructions and explanations were clear. All panelists *strongly agreed* that they were prepared to make their standard-setting judgments. All panelists *strongly agreed* or *agreed* that the standard-setting process was easy to follow.

All panelists reported that the description of the just-qualified candidate was *very influential* in guiding their standard-setting judgments. All of the panelists reported that between-round discussions were at least *somewhat influential* in guiding their judgments. Three of the seven panelists indicated that their own professional experience was *very influential* in guiding their judgments.

All of the panelists indicated they were at least *somewhat comfortable* with the passing score they recommended; five of the panelist were *very comfortable* with the recommended passing score. All of the panelists indicated the recommended passing score was *about right*.

Summary

To support the decision-making process for education agencies establishing a passing score (cut score) for the *Praxis* Russian: World Language test, research staff from ETS designed and conducted a multistate standard-setting study.

ETS provides a recommended passing score from the multistate standard-setting study to help education agencies determine an appropriate operational passing score. For the *Praxis* Russian: World Language test, the recommended passing score is 39 out of a possible 98 raw-score points. The scale score associated with a raw score of 39 is 130 on a 100–200 scale.

References

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17*, 59–88.

Brennan, R. L. (2002, October). Estimated standard error of a mean when there are only two observations (Center for Advanced Studies in Measurement and Assessment Technical Note Number 1). *Iowa City: University of Iowa*.

Cizek, G. J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

ETS. (in press). *The Praxis Series®: The Praxis Study Companion: Russian: World Language (5671)*. Princeton, NJ: Author.

Geisinger, K. F., & McCormick, C. M. (2010), Adopting cut scores: post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*, 38–44.

Guilford, J. P. (1965). Fundamental statistics in psychology and education (4th ed.). New York: McGraw-Hill, pp. 158-159.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., & Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41-55.

High, R. (2000). Dealing with 'outliers': How to maintain your data's integrity. *University of Oregon Computing News*, *15*(3), 14-16. Retrieved from http://hdl.handle.net/1794/3129

Lord. F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21*, 239-283.

MacCann, R.G., & Stanley, G. (2004). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical Assessment: Research and Evaluation*, *9*. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=5

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

Tannenbaum, R. J. (2011). Setting standards on *The Praxis Series*™ tests: A multistate approach. *R&D Connections*, *17*, 1-9.

Tannenbaum, R. J. (2012). A multistate approach to setting standards: An application to teacher licensure tests. *CLEAR Exam Review*, *23*(1), 18-24.

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.

Walker, H. M., & Lev, J. (1953). *Statistical inference*. New York: Holt Rinehart and Winston (pp. 414-415).

Appendix A: Panelists' Names & Affiliations

Participating Panelists With Affiliation and State

Panelist Name	**Panelists' Affiliation and State Abbreviation**
Nadja Berkovich	University of Arkansas (AR)
Annalisa Czeczulin	Goucher College (MD)
Natalia Howard	Overlake Elementary School (UT)
Albina Parks	Baltimore International Academy (MD)
Julia Revok	Baltimore International Academy (MD)
Irina Six	University of Kansas (KS)
Yana Willey	Baltimore International Academy (MD)

Note. An additional educator (from Indiana) was in attendance during half of the first day but was unable to continue after test familiarization. This educator is not described in this report as a part of the panel because they did not participate in any of the collaborative work that contributed to the passing score recommendation—such as the creation of the just-qualified candidate description.

Appendix B: Agenda

Praxis® Russian: World Language (5671)

Standard-Setting Study

Day 1 Agenda

Welcome and Introduction

Overview of Standard Setting and the *Praxis* Russian: World Language Test

Review the *Praxis* Russian: World Language Test

AM Break

Discuss the *Praxis* Russian: World Language Test

Lunch

Define the Knowledge/Skills of a Just-Qualified Candidate (small group drafts)

PM Break

Define the Knowledge/Skills of a Just-Qualified Candidate (small group drafts) *(continued)*

Collect Materials; End of Day 1

Praxis® Russian: World Language (5671)

Standard-Setting Study

Day 2 Agenda

Overview of Day 2

Define the Knowledge/Skills of a Just-Qualified Candidate (whole-group consensus)

AM Break

Standard-Setting Training in the Modified Angoff Method

Practice Round – Independent Judgments

Lunch

Practice Round –Discussion

Round 1 Standard Setting Judgments for Selected-Response Items

PM Break

Round 1 Standard Setting Judgments for Selected-Response Items *(continued)*

Collect Materials; End of Day 2

Praxis® Russian: World Language (5671)

Standard-Setting Study

Day 3 Agenda

Overview of Day 3

Honoraria Presentation

Standard Setting Training in the Extended Angoff Method

AM Break

Practice Round –Discussion

Round 1 Standard Setting Judgments for Constructed-Response Items

AM Break

Round 1 Feedback and Round 2 Judgments

Lunch

Round 1 Feedback and Round 2 Judgments *(continued)*

Feedback on Round 2 Recommended Passing Score

Complete Final Evaluation

Collect Materials; End of Study

Appendix C: Just-Qualified Candidate Description

Description of the Just-qualified candidate[7]

A just-qualified candidate…

Listening, Reading, and Cultural Knowledge

1. Has an intermediate-high level of understanding of spoken and written Russian
2. Uses basic reading strategies such as word analysis, inference, and context clues with authentic (appropriate for intermediate high level) texts
3. Has an intermediate-high ability to understand a wide range of Russian speakers (e.g., native Russian speaker sympathetic to L2 learners)
4. Has an understanding of intermediate-level grammar, syntactical relationships, and the interaction of tense and aspect
5. Comprehends a commonly-used Russian vocabulary encompassing a variety of practical topics, including basic idiomatic expressions
6. Grasps the main idea, most subordinate ideas, and some details in authentic aural and written communication
7. Recognizes various registers and formal/informal voice in authentic aural and written communication
8. Is familiar with significant current, historical, ethnic/linguistic, and religious events, people, places in Russia

Writing and Speaking

9. Is easily comprehensible to a native Russian speaker sympathetic to L2 learners, through the use of commonly-used Russian vocabulary, varied grammatical and syntactical forms, and circumlocution as necessary in writing and speaking
10. Is comprehensible in articulation, pronunciation, and fluency to a sympathetic L1 speaker
11. Can express themselves at an intermediate-high level in an organized, cohesive manner using Russian vocabulary that encompasses a variety of simple practical topics
12. Demonstrates an intermediate command of mechanics and conventions in speaking and writing
13. Employs formal/informal registers for various purposes in spoken and written communication

[7] Description of the just-qualified candidate focuses on the knowledge/skills that differentiate a *just* from a *not quite* qualified candidate.

Appendix D: Technical Notes

Standard Error of Judgment (SEJ)

The standard error of judgment (SEJ) is one way of estimating the reliability or consistency of a panel's standard-setting judgments. It indicates how likely it would be for several other panels of educators similar in makeup, experience, and standard-setting training to the current panel to recommend the same threshold score on the same form of the assessment. The SEJ assumes that panelists are randomly selected and that standard-setting judgments are independent. It is seldom the case that panelists are randomly sampled, and only the first round of judgments may be considered independent. The SEJ, therefore, likely underestimates the uncertainty of threshold scores (Tannenbaum & Katz, 2013).

The SEJ is calculated by dividing the standard deviation of the panelists' judgments (*SD*) by the square root of the number of panelists (*n*). The result serves as an estimate of the standard error of the mean (Brennan, 2002).

$$\text{SEJ} = \text{SD}/\sqrt{n}$$

Outlier Analysis

An analysis of the data is conducted per panel. Judgments that are above or below 1.5 times the interquartile range for that panel are identified as outliers (High, 2000). ETS makes recommendations on the removal of specific outliers based on the observations of the panel facilitator. The panel facilitator reports whether or not the specified panelist was faithfully participating in the standard-setting process. The decision to accept the panel recommendation with or without the outlier data is solely at the discretion of the state.

Estimated Conditional Standard Error of Measurement (CSEM)

The estimated conditional standard error of measurement (*CSEM*) for a test consisting of both selected-response and constructed-response questions is equal to the square root of the sum of the squared CSEM for selected-response items ($CSEM_{SR}$) and the squared CSEM for constructed response items ($CSEM_{CR}$).

$$CSEM = \sqrt{(CSEM_{SR})^2 + (CSEM_{CR})^2}$$

Where $CSEM_{SR}$ is computed from the study value (*SV*) of the recommended passing score and the number of selected-response items (*n*) on the test (see Lord, 1984):

$$CSEM_{SR} = \sqrt{(SV)(n - SV)/(n - 1)}$$

and $CSEM_{CR}$ is computed as

$$CSEM_{CR} = SD\sqrt{(1 - r)}$$

Where the internal consistency reliability index, *r*, is set equal to .75 (a lower bound estimate) and the standard deviation (*SD*) is estimated as

$$SD = ([.95][MAX] - MIN)/6$$

MAX equals the maximum possible raw score for the constructed-responses items. *MIN* equals the rounded value of ([.05][*MAX*]).

Appendix E: Final Evaluation Results

Table E1: Final Evaluation: Process Questions

Likert Statement	Strongly agree *N*	Strongly agree %	Agree *N*	Agree %	Disagree *N*	Disagree %	Strongly disagree *N*	Strongly disagree %
I understood the purpose of this study.	3	43	4	57	0	0	0	0
The instructions and explanations provided by the facilitators were clear.	4	57	3	43	0	0	0	0
The training in the standard-setting method was adequate to give me the information I needed to complete my assignment.	7	100	0	0	0	0	0	0
The explanation of how the recommended passing score is computed was clear.	6	86	1	14	0	0	0	0
The opportunity for feedback and discussion for round 2 judgments was helpful.	7	100	0	0	0	0	0	0
The process of making the standard-setting judgments was easy to follow.	4	57	3	43	0	0	0	0

Table E2: Final Evaluation: Standard-Setting Process

	Too much time *N*	Too much time %	About the right amount of time *N*	About the right amount of time %	Too little time *N*	Too little time %
Small group JQC drafts	0	0	7	100	0	0
Whole group JQC consensus	0	0	7	100	0	0
Training and practice for making standard-setting judgments	0	0	7	100	0	0
Round 1 judgments (independent)	1	14	5	71	1	14
Round 2 judgments (with discussion)	0	0	7	100	0	0

Table E3: Final Evaluation: Influences in Standard-Setting Judgments

How influential was each of the following factors in guiding your standard-setting judgments?	Very influential *N*	Very influential %	Somewhat influential *N*	Somewhat influential %	Not influential *N*	Not influential %
The description of the just-qualified candidate	7	100	0	0	0	0
The round 2 discussion	6	86	1	14	0	0
The knowledge/skills required to answer each test item	4	57	3	43	0	0
The passing scores of other panel members	5	71	1	14	1	14
My own professional experience	3	43	4	57	0	0

Table E4: Final Evaluation: Comfort with the Panel's Recommendation

Question	Very comfort-able *N*	Very comfort-able %	Somewhat comfort-able *N*	Somewhat comfort-able %	Somewhat uncom-fortable *N*	Somewhat uncom-fortable %	Very uncom-fortable *N*	Very uncom-fortable %
Considering the process you followed, how comfortable are you with the panel's recommended cut score?	5	71	2	29	0	0	0	0

Table E5: Final Evaluation: Opinion of the Final Recommendation

Statement	Too low *N*	Too low %	About right *N*	About right %	Too high *N*	Too high %
Overall, the recommended passing score is:	0	0	7	100	0	0