Measuring the Power of Learning.™

Research Memorandum
ETS RM–15-11

Recommending a Passing Score for the *Praxis*® Performance Assessment for Teachers (PPAT)

Clyde M. Reese

Richard J. Tannenbaum

October 2015

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Recommending a Passing Score for the *Praxis*® Performance Assessment for Teachers (PPAT)

Clyde M. Reese and Richard J. Tannenbaum
Educational Testing Service, Princeton, New Jersey

October 2015

Corresponding author: C. Reese, E-mail: CReese@ets.org

Suggested citation: Reese, C. M., & Tannenbaum, R. J. (2015). *Recommending a passing score for the Praxis® Performance Assessment for Teachers (PPAT)* (Research Memorandum No. RM-15-11). Princeton, NJ: Educational Testing Service.

Action Editor: Heather Buzick

Reviewers: Geoffrey Phelps and Priya Kannan

Abstract

A standard-setting workshop was conducted with 12 educators who mentor or supervise preservice (or student teacher) candidates to recommend a passing score for the *Praxis*® Performance Assessment for Teachers (PPAT). The multiple-task assessment requires candidates to submit written responses and supporting instructional materials and student work (i.e., artifacts). The last task, Task 4, also includes submission of a video of the candidate's teaching. A variation on a multiple-round extended Angoff method was applied. In this approach, for each step within a task, a panelist decided on the score value that would most likely be earned by a just qualified candidate (Round 1). Step-level judgments were then summed to calculate task-level scores for each panelist and panelists were able to adjust their judgments at the task level (Round 2). Finally, task-level judgments were summed to calculate a PPAT score for each panelist and panelists were able to adjust their overall scores (Round 3). The recommended passing score for the overall PPAT is 40 out of a possible 60 points. Procedural and internal sources of evidence support the reasonableness of the recommended passing scores.

Key words: *Praxis*®, PPAT, standard setting, cut scores, passing scores

The impact of teachers in the lives of students is widely accepted (Harris & Rutledge, 2010) and the importance of teacher quality in student achievement is well established (e.g., Ferguson, 1998; Goldhaber, 2002; Rivkin, Hanushek, & Kain, 2005). While knowledge of the content area is an obvious prerequisite, teaching behavior also is critical when examining teacher quality (Ball & Hill, 2008). Efforts to assist educator preparation programs and state teacher licensure agencies to improve teacher quality can start with examining teaching quality at the point of entry into the profession and the licensure and certification processes that are intended to safeguard the public. Licensure assessments, as part of a larger licensure process, can include teaching behaviors as well as content knowledge—both subject matter and pedagogical.

The *Praxis*® Performance Assessment for Teachers (PPAT) is a multiple-task, authentic performance assessment completed during a candidate's preservice, or student teaching, placement. The PPAT measures a candidate's ability to gauge his or her students' learning needs, interact effectively with students, design and implement lessons with well-articulated learning goals, and design and use assessments to make data-driven decisions to inform teaching and learning. A multiple-round standard-setting study was conducted in June 2015 to recommend a passing score for the PPAT. This report documents the standard-setting procedures and results of the study.

Standard Setting

Licensure assessments, like the PPAT, are intended to be mechanisms that provide the public with evidence that candidates passing the assessment and entering the field have demonstrated a particular level of knowledge and skills (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Establishing the performance standard—the minimum assessment score that differentiates between *just qualified* and *not quite qualified*—is the function of standard setting (Tannenbaum, 2011). For licensure assessments, where assessment scores are used in part to award or deny a license to practice, standard setting is critical to the validity of the test score interpretation and use (Bejar, Braun, & Tannenbaum, 2007; Kane, 2006; Margolis & Clauser, 2014; Tannenbaum & Kannan, 2015).

Educational Testing Service (ETS), as the publisher of the PPAT, provides a recommended passing score from a 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. Standard setting is a judgment-based process; there is not an empirically correct passing score (O'Neill, Buckendahl, Plake, & Taylor, 2007). The value of the recommended passing score rests on the appropriateness of the study design given the structure and content of the test and the quality of the implementation of that design (Tannenbaum & Cho, 2014). Each state may want to consider the recommended passing score but also other sources of information when setting the final 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 it meeting the state's needs.

Overview of the PPAT

The PPAT is a multiple-task, authentic performance assessment designed for teacher candidates to complete during their preservice, or student teaching, placement. Development of the PPAT by ETS began in 2013, field testing occurred in 2014–15, and the operational launch is scheduled for fall 2015. The assessment is composed of four tasks:

- Task 1: Knowledge of Students and the Learning Environment
- Task 2: Assessment and Data Collection to Measure and Inform Student Learning
- Task 3: Designing Instruction for Student Learning
- Task 4: Implementing and Analyzing Instruction to Promote Student Learning

All tasks include written responses and supporting instructional materials and student work (i.e., artifacts). Task 4 also includes submission of a video of the candidate's teaching.

The content of the PPAT is aligned with Interstate Teacher Assessment and Support Consortium (InTASC) *Model Core Teaching Standards* (CCSSO, 2013). Task 1 is formative and candidates will work with their preparation programs to receive feedback on this task. Tasks 2, 3, and 4 are summative; scores for these tasks, as well as the weighted sum of the three task scores, will be reported. (The standard-setting study provides a recommended passing score for the overall PPAT score, which is the weighted sum of scores on Tasks 2, 3, and 4.)

Each task is composed of steps: Task 1 includes two steps, Task 2 includes three steps, and Tasks 3 and 4 include four steps each. Task 1 is formative and scored by a candidate's supervising faculty. Tasks 2, 3, and 4 are summative and centrally scored. Each step within a task is scored using a step-specific, 4-point rubric. The maximum score for Task 2 is 12 points (the range is 3–12) and for Task 3 is 16 points (the range is 4–16). The score for Task 4 is doubled; therefore, the maximum score is 32 (the range is 8–32). For the overall PPAT, the maximum score is 60 (the range is 15–60).

Panelists

The multistate standard-setting panel was composed of 12 educators from eight states (Delaware, Hawaii, Iowa, North Carolina, North Dakota, New Jersey, Pennsylvania, and West Virginia). The number of panelists fell within an acceptable range, from 10 to 15 panelists (Hurtz & Hertz, 1999; Raymond & Reid, 2001). All the educators are involved with the preparation and supervision of prospective teachers. The majority of panelists (nine of the 12 panelists) were college faculty or associated with a teacher preparation program; the remaining three panelists worked in K–12 school settings. All the panelists reported mentoring or supervising preservice, or student, teachers in the past five years. Most (10 of 12 panelists) had at least 15 years' experience mentoring or supervising preservice teachers (see Table 1).

Table 1. Panelists Background

Characteristic	*N*	%
Current position		
K–12 teacher	2	17
Administrator	1	8
College faculty	9	75
Gender		
Female	8	67
Male	4	33
Race		
White	4	33
Black or African American	5	42
Hispanic or Latino	1	8
Asian or Asian American	2	17
Mentored or supervised preservice teachers in the past 5 years		
Yes	12	100
No	0	0
Experience mentoring or supervising preservice teachers		
3 years or less	0	0
4–9 years	2	17
10–14 years	0	0
15 years or more	10	83
No experience	0	0

Procedures

A variation on a multiple-round extended Angoff method (Plake & Cizek, 2012; Tannenbaum & Katz, 2013) was used for the PPAT. In this approach, for each step within a task, a panelist decided on the score value that would most likely be earned by a just-qualified candidate (JCQ; Round 1). Step-level judgments were then summed to calculate task-level scores for each panelist and panelists were able to adjust their judgments at the task-level (Round 2). Finally, task-level judgments were summed to calculate a PPAT score for each panelist and panelists were able to adjust their overall scores (Round 3).

Reviewing the PPAT

Approximately 2 weeks prior to the study, panelists were provided available PPAT materials, including the tasks, scoring rubrics, and guidelines for preparing and submitting supporting artifacts. The materials panelists reviewed were the same materials provided to candidates. Panelists were asked to take notes on tasks or steps within tasks, focusing on what is being measured and the challenge the task poses for preservice teachers.

At the beginning of the study, ETS performance assessment specialists described the development of the tasks and the administration of the assessment. Then, the structure of each task—prompts, candidate's written response, artifacts, and scoring rubrics—were described for the panel. The whole-group discussion focused on what knowledge/skills are being measured, how candidates respond to the tasks and what supporting artifacts are expected, and what evidence is being valued during scoring.

Defining the Just-Qualified Candidate (JQC)

Following the review of the PPAT, panelists engaged in the process described below to describe the JQC. The JQC 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 (Tannenbaum & Katz, 2013). The emphasis on minimally sufficient knowledge and skills when describing the JQC is purposeful. This is because the passing score, which is the numeric equivalent of the performance expectations described in the JQC, is intended to be the lowest acceptable score that denotes entrance into the passing category. The panelists drew upon their experience with having reviewed the PPAT and their own experience mentoring or supervising preservice teachers when discussing the JQC description.

During a prior alignment study (Reese, Tannenbaum, & Kuku, 2015), a separate panel of subject-matter experts identified the InTASC standards performance indicators being measured by the PPAT. The results of the alignment study served as the preliminary JCQ description. The standard-setting panelists independently reviewed the 38 knowledge/skill statements identified by the alignment study and rated if each statement was more than would be expected of a JQC, less than would be expected, or about right. Ratings were summarized and each statement was discussed by the whole group. Panelists offered qualifiers to some statements to better describe the performance of a just-qualified preservice teacher, and panelists were encouraged to take notes on the JQC description for future reference. For 29 of the 38 statements, half or more of the panelists rated the statement as *about right* for a JQC. For another five statements (Statements 13, 18, 21, 30, and 37), half or more of the panelists rated the statement as *more than* would be expected of a JQC. For these statements, panelists discussed how a JQC would have an awareness of appropriate approaches or responses but their demonstration may be restricted to common occurrences (e.g., Statements 13 and 18) or may be limited in depth or experience (e.g., Statements 21, 30, and 37 dealing with assessments/data). Panelists were instructed to make notes on their printed copy of the statements that added qualifiers (e.g., “basic awareness of” or “common misconceptions”) to bring the statement in line with agreed-upon expectations for a JQC. The remaining four statements received mixed rating; however, after discussion the panel agreed they were *about right* for a JQC. All 38 knowledge/skill statements that formed the JQC description are included in the appendix. Each panelist referred to his or her annotated JQC description during the study that included notes from the prior discussion (i.e., qualifiers for some statements).

Panelists’ Judgments

The following steps were followed for each task. The panel completed Rounds 1 and 2 for a task before moving to the next task. Round 3 was completed after Rounds 1 and 2 were completed for all three tasks. The judgment process started with Task 2 and was repeated for Tasks 3 and 4. The committee did not consider Task 1. Figure 1 summarizes the standard-setting process.

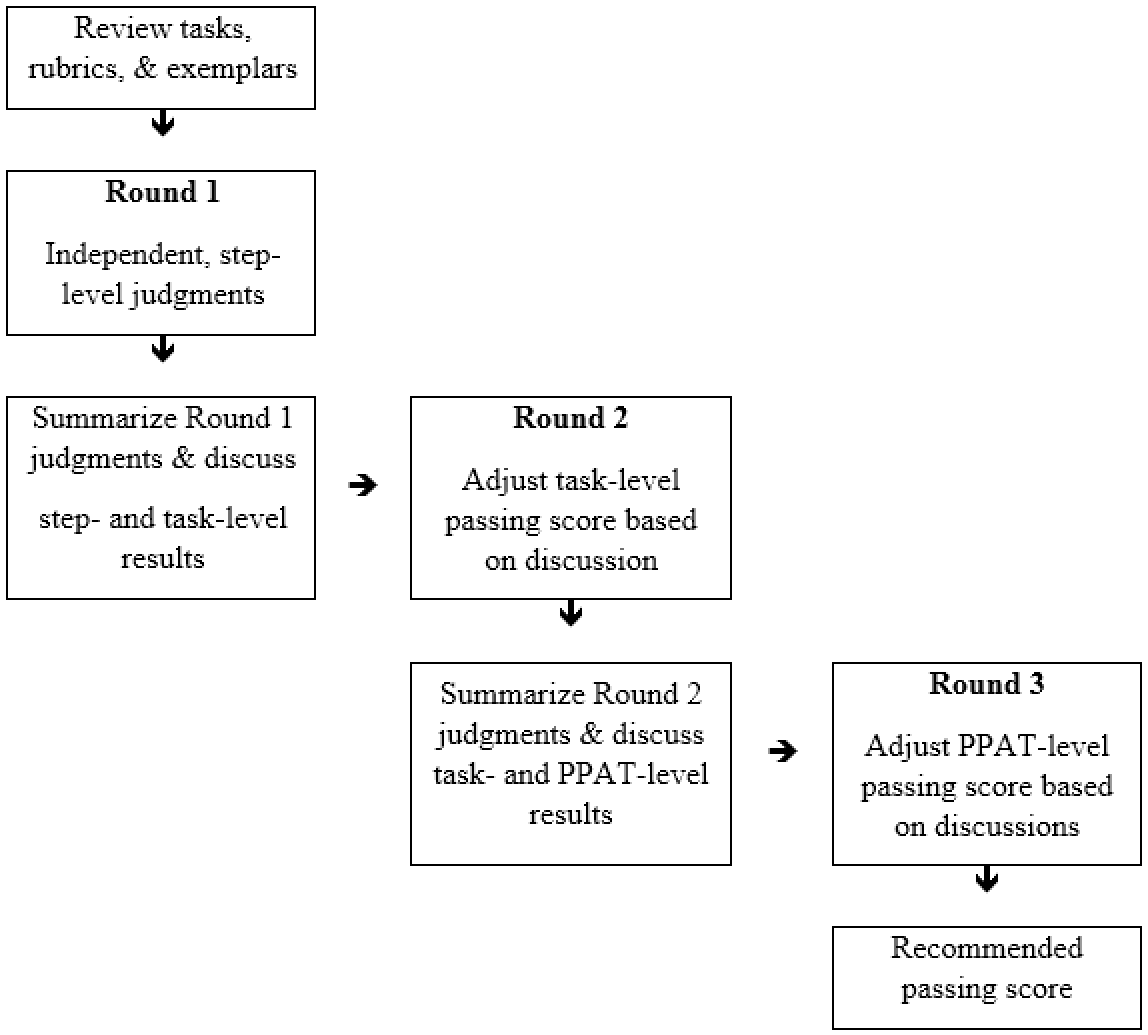
Review tasks, rubrics, & exemplars
Round 1
Independent, step-level judgments
Summarize Round 1 judgments & discuss step- and task-level results
Round 2
Adjust task-level passing score based on discussion
Summarize Round 2 judgments & discuss task- and PPAT-level results
Round 3
Adjust PPAT-level passing score based on discussions
Recommended passing score

Figure 1. PPAT standard-setting process.

Review PPAT materials. An ETS performance assessment specialist conducted an in-depth review of the task. The review focused on the specific components of each step, how the artifacts support a candidate's responses, and the step-specific rubrics. The step-level scoring process and how step-level scores are combined to produce the task-level score were highlighted. The panel also reviewed exemplars of each score point for each step within a task.

Round 1 judgments. The panelists reviewed the task, the rubrics, and exemplars. Then the panelists independently judged, for each step within the task, the score (1, 2, 3, 4) a JQC would likely receive. Panelists were allowed to assign a judgment between rubric points;[1]

therefore, the judgment scale was 1, 1.5, 2, 2.5, 3, 3.5, and 4. The task-level result of Round 1 is the simple sum of the likely scores for each step.

Round 2 judgments. Round 1 judgments were collected and summarized. Frequency distributions of the step- and task-level judgments were presented with the average highlighted. Table 2 presents a sample of the Round 1 results (for Task 2) that were shared with the panel. Discussions first focused on the step-level judgments and then turned to the task-level. The panelists were asked if their task-level score from Round 1 (the sum of the step-level judgments) reflected the likely performance of a JQC, considering the various patterns of step scores that may result in a task score, or if their task-level score should be adjusted. Following the discussion, the panelists provided a task-level Round 2 judgment. Panelists could maintain their Round 1 judgment or adjust up or down based on the discussion.

Table 2. Sample Round 1 Feedback: Task 2

Score	Step 1	Step 2	Step 3	Task score
1	0	0	0	
1.5	0	0	0	
2	1	5	3	
2.5	8	7	9	
3	3	0	0	
3.5	0	0	0	
5	0	0	0	
Mean	2.6	2.3	2.4	7.3
Median	2.5	2.5	2.5	7.0
Minimum	2.0	2.0	2.0	6.5
Maximum	3.0	2.5	2.5	8.0
SD	0.29	0.26	0.23	0.45

Round 3 judgments. Following Rounds 1 and 2 for the three tasks, frequency distributions of the task- and assessment-level judgments were presented with the average highlighted. Discussions first focused on the task-level judgments and then turned to the recommended passing score for the assessment. The panelists were asked if their assessment-level score from Round 2 (the weighted sum[2] of the task-level judgments) reflected the likely performance of a JQC, considering the various patterns of task scores that may result in a PPAT score, or if their assessment-level score should be adjusted. Following the discussion, the panelists provided an assessment-level Round 3 judgment. Panelists could maintain their Round 2 judgment or adjust up or down based on the discussion.

Final Evaluations

The panelists completed an evaluation form at the conclusion of the study addressing the quality of the standard-setting implementation and their acceptance of the recommended passing score. The responses to the evaluation provide evidence of the validity of the standard-setting process and the reasonableness of the passing score (Hambleton & Pitoniak, 2006; Kane, 2001).

Results

Recommended Passing Score

Standard-setting judgments were collected first at the step-level (Round 1) and then adjusted at the task-level (Round 2) and assessment-level (Round 3). Table 3 summarizes the task-level judgments after Rounds 1 and 2 and the assessment-level judgments after Round 3 for each of the 12 panelists. Task 2 is composed of three steps; Tasks 3 and 4 are composed of four steps each. The task-level Round 1 results were calculated by summing the step-level judgments. The mean and median task-level results differed by less than half a point for each task between Rounds 1 and 2; the standard deviations decreased for each task.

The mean and median of the Round 3 judgments differed by 0.1 points on a 45-point scale (15–60). Rounding rules would result in the mean score (40.1) being translated to a recommended cut score of 40.5. The median score (40.0) better reflects the distribution of panelists' Round 3 results. A recommended passing score of 40.5 would be higher than nine of the 12 panelist recommendations. A recommended passing score of 40.0 was the Round 3 score for seven of the 12 panelists. Therefore, the panel's recommended passing score for the PPAT is the median of the weighted sum of the three task scores following Round 3.

Sources of Evidence Supporting the Passing Score

Standard setting is a judgment-based process that relies on the considered judgments of subject-matter experts. The resulting passing score, when applied in a high-stakes situation such as initial teacher licensure, carries considerable weight. The confidence the public places on the recommended passing score is bolstered by procedural evidence and internal evidence (Kane, 1994, 2001). Procedural evidence refers to the quality of the standard-setting study and internal evidence refers to the likelihood of replicating the recommended passing score. Results addressing these two sources of evidence are presented below.

Table 3. Passing-Score Recommendation by Round and Task: Rounds 1 and 2

	Round 1[a]			Round 2[a]			Round 3
Panelist	Task 2[b]	Task 3[c]	Task 4[d]	Task 2[b]	Task 3[c]	Task 4[d]	Overall[e]
1	7.5	10.5	10.5	7	10.5	10.5	39
2	7	11	11	8	10.5	11	40.5
3	7	14	11	7	12	11	41
4	8	10	11.5	7.5	11	11	40
5	6.5	13.5	10	7	12.5	10.5	40
6	8	10.5	10.5	7.5	11.5	10.5	40
7	7.5	10.5	10.5	7.5	11	10.5	39.5
8	7.5	10.5	10.5	7.5	11	10.5	40
9	7	11	11	7	11	10.5	40
10	7	9.5	11	7	10.5	11	40
11	7	11	11.5	7	11	11	40
12	7	11.5	11.5	7.5	12.5	11	41
Mean	7.3	11.1	10.9	7.3	11.3	10.8	40.1
Median	7.0	10.8	11.0	7.3	11.0	10.8	40.0
Minimum	6.5	9.5	10.0	7.0	10.5	10.5	39.0
Maximum	8.0	14.0	11.5	8.0	12.5	11.0	41.0
SD.	0.45	1.33	0.48	0.33	0.72	0.26	0.56
SEJ$_{\text{Median}}$	0.16	0.48	0.17	0.12	0.26	0.09	0.20

[a] For Rounds 1 and 2, recommended scores for Task 4 are unweighted. [b] Possible candidate scores for Task 2 range from 3 to 12. [c] Possible candidate scores for Task 3 range from 4 to 16. [d] Possible candidate scores for Task 4 range from 4 to 16. [e] Recommended scores for Task 4 are weighted to calculate the overall score; possible candidate scores range from 15 to 60.

Procedural evidence. Procedural evidence often comes from panelists' responses to the training and end-of-study evaluations (Cizek 2012; Cizek & Bunch, 2007). Following training for each of the three rounds of judgments, all 12 panelists verified that they understood the process and confirmed their readiness to proceed.

Following the completion of the study, the panelists completed a poststudy evaluation (see Table 4). The panelists were asked (a) if they understood the purpose of the study, (b) if instructions and explanation provided were clear, (c) if they were adequately trained, and (d) if the process was easy to follow. All the panelists *strongly agreed* that they understood the purpose and that the instructions and explanations were clear. All the panelists *agreed* or *strongly agreed* that they were adequately trained and that the process was easy to follow.

Table 4. Poststudy Evaluation

Statement	*N* (%)	*N* (%)	*N* (%)	*N* (%)
	Strongly agree	Agree	Disagree	Strongly disagree
I understood the purpose of this study.	12 (100%)	0 (0%)	0 (0%)	0 (0%)
The instructions and explanations were clear.	12 (100%)	0 (0%)	0 (0%)	0 (0%)
The training in the standard setting method was adequate to give me the information I needed to complete my assignment.	11 (92%)	1 (8%)	0 (0%)	0 (0%)
I understood the PPAT tasks/steps well enough to make my judgments.	12 (100%)	0 (0%)	0 (0%)	0 (0%)
I understood the PPAT rubrics well enough to make my judgments.	12 (100%)	0 (0%)	0 (0%)	0 (0%)
The exemplars were helpful in describing levels of performance.	10 (83%)	2 (17%)	0 (0%)	0 (0%)
The explanation of how the recommended cut score is computed was clear.	11 (92%)	1 (8%)	0 (0%)	0 (0%)
The opportunity for feedback and discussion between rounds was helpful.	12 (100%)	0 (0%)	0 (0%)	0 (0%)
The process of making the standard setting judgments was easy to follow.	11 (92%)	1 (8%)	0 (0%)	0 (0%)
	Very comfortable	Somewhat comfortable	Somewhat uncomfortable	Very uncomfortable
Overall, how comfortable are you with the panel's recommended passing score[a]?	11 (92%)	1 (8%)	0 (0%)	0 (0%)
	Too low	About right	Too high	
Overall, the panel's recommended passing score[a] is:	0 (0%)	12 (100%)	0 (0%)	

[a] Panelists provided their confidence judgments for the passing score based on the panel mean (40.5) rather than the poststudy recommended passing score based on the panel median (40.0).

The panelists also were asked if they understood the PPAT tasks/steps and rubrics well enough to make their judgments and if the exemplars were helpful. All the panelists *strongly agreed* that they understood the PPAT tasks/steps and rubrics. All the panelists *agreed* or *strongly agreed* that the exemplars were helpful in describing levels of performance.

In addition to the panelists' evaluation of the standard-setting process, they also were shown the panel's recommended passing score[3] and asked (a) how comfortable they were with the recommended passing score and (b) if they thought the score was too high, too low, or about right. All but one of the panelists were *very comfortable* with the passing score they recommended; the remaining panelist indicated he was *somewhat comfortable*. All the panelists indicated that recommended passing score was *about right.*

Internal evidence. Internal evidence (consistency) addresses the likelihood of replicating the recommended passing score. For a single panel standard-setting study, an approximation of replicability is provided by the standard error associated with the recommended passing scores (Cizek & Bunch, 2007; Kaftandjieva, 2010). This standard error of judgment (SEJ) is an index of the extent to which the passing score would vary if the study were repeated with different panels of educators (Zieky, Perie, & Livingston, 2008). The smaller the value is, the less likely it is that other panels would recommend a significantly different passing score. A general guideline for interpreting the SEJ is its size relative to the standard error of measurement (SEM) of the test. According to Cohen, Kane, and Crooks (1999), an SEJ less than one-half of the SEM is considered reasonable. An estimate of the SEM for the PPAT (total score calculated as the weighted sum of Tasks 2, 3, and 4) from a field test of nearly 200 preservice teachers was 4.35. The SEJ for the median of the panelists' judgments[4] from the study is 0.20, well below half the value of the SEM.

Summary

The PPAT was designed to be a component of a state's initial teacher licensure system. In this score-use context, each state's department of education, board of education, or designated educator licensure board is responsible for establishing the operational passing score in accordance with applicable regulations. A standard-setting study was conducted with 12 educators who monitor or supervise preservice teacher. The recommended passing score, the median of the judgments for the panel, is 40 out of a possible 60 points for the overall PPAT (weighted sum of Tasks 2, 3, and 4 scores).

Although both procedural and internal sources of evidence support the reasonableness of the recommended passing score, the final responsibility for establishing the passing score rests with the state-level entity authorized to award initial teacher licenses. Establishing a performance standard (i.e., passing score) on a licensure assessment like PPAT is comparable to establishing or forming a policy, where decisions are neither right or wrong (Kane, 2001). Each state may want to consider the recommended passing score and also other sources of information when setting the final 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.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ball, D. L., & Hill, H. C. (2008). Measuring teacher quality in practice. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality* (pp. 80–98). Thousand Oaks, CA: Sage.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1–30). Maple Grove, MN: JAMA Press.

CCSSO. (2013). *InTASC model core teaching standards and learning progressions for teachers 1.0*. Retrieved from http://programs.ccsso.org/content/pdfs/corestrd.pdf

Cizek, G. J. (2012). The forms and functions of evaluations in the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 163–178). New York, NY: Routledge.

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

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*(4), 343–366.

Ferguson, R. F. (1998). Can schools narrow the Black-White test score gap? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 318–374). Washington, DC: Brookings Institution.

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. http://dx.doi.org/10.1111/i.1745-3992.2009.00168.x

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, *2*(1), 50–55.

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

Harris, D. N., & Rutledge, S. A. (2010). Models and predictors of teacher effectiveness: A comparison of research about teaching and other occupations. *Teachers College Record, 112*(3), 914–960.

Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability study. *Educational and Psychological Measurement, 59,* 885–897.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem, The Netherlands: CITO.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425–462. http://dx.doi.org/10.3102/00346543064003425

Kane, M. T. (2001). So much remains the same: Conceptions and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.

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

Margolis, M. J., & Clauser, B. E. (2014). The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. *Educational Measurement: Issues and Practice, 33,* 15–22. http://dx.doi.org/10.1111/emip.12025

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly, 4,* 295–317. http://dx.doi.org/10.1080/15434300701533562

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27,* 15–29. http://dx.doi.org/10.1111/j.1745-3992.2008.00135.x

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance* *standards: Foundations, methods, and innovations* (2nd ed., pp. 181–199). New York, NY: Routledge.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Erlbaum.

Reese, C. M., Tannenbaum, R. J., & Kuku, B. (2015). Alignment between the Praxis Performance Assessment for Teachers (PPAT) and the Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards (Research Memorandum No. RM-15-10). Princeton, NJ: Educational Testing Service.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73(2), 417–458.

Tannenbaum, R. J. (2011). Standard setting. In J. W. Collins & N. P. O’Brien (Eds.), Greenwood dictionary of education (2nd ed.). Santa Barbara, CA: ABC-CLIO

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. Language Assessment Quarterly, 11, 233–249. http://dx.doi.org/10.1080/15434303.2013.869815

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of Angoff-based judgments: Are item judgments and passing scores replicable across different panels of experts? Educational Assessment, 20(1), 66–78. http://dx.doi.org/10.1080/10627197.2015.997619

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.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). Cutscores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Appendix. Description of the Just Qualified Candidate (JQC)

JQC Description for the PPAT Standard Setting Study

1. Drawing on her/his understanding of child and adolescent development, the teacher observes learners, noting changes and patterns in learners across areas of development, and seeks resources, including from families and colleagues, to adjust teaching.

2. The teacher actively seeks out information about learner interests in order to engage learners in developmentally appropriate learning experiences.

3. Drawing upon her/his understanding of second language acquisition, exceptional needs, and learners' background knowledge, the teacher observes individual and groups of learners to identify specific needs and responds with individualized support, flexible grouping, and varied learning experiences.

4. Recognizing how diverse learners process information and develop skills, the teacher incorporates multiple approaches to learning that engage a range of learner preferences.

5. Using information on learners' language proficiency levels, the teacher incorporates tools of language development into planning and instruction, including strategies for making content and academic language accessible to linguistically diverse learners.

6. The teacher includes multiple perspectives in the presentation and discussion of content that include each learner's personal, family, community, and cultural experiences and norms.

7. The teacher applies interventions, modifications, and accommodations based on IEPs, IFSPs, 504s and other legal requirements, seeking advice and support from specialized support staff and families.

8. The teacher follows a process, designated by a school or district, for identifying and addressing learner needs (e.g., Response to Intervention) and documents learner progress.

9. The teacher communicates verbally and nonverbally in ways that demonstrate respect for each learner.
10. The teacher is a responsive and supportive listener, seeing the cultural backgrounds and differing perspectives learners bring as assets and resources in the learning environment.
11. The teacher manages the learning environment, organizing, allocating and coordinating resources (e.g., time, space, materials) to promote learner engagement and minimize loss of instructional time.
12. The teacher accurately and effectively communicates concepts, processes and knowledge in the discipline, and uses vocabulary and academic language that is clear, correct and appropriate for learners.
13. The teacher draws upon his/her initial knowledge of common misconceptions in the content area, uses available resources to address them, and consults with colleagues on how to anticipate learner's need for explanations and experiences that create accurate understanding in the content area.
14. The teacher engages learners in applying methods of inquiry used in the discipline.
15. The teacher links new concepts to familiar concepts and helps learners see them in connection to their prior experiences.
16. The teacher models and provides opportunities for learners to understand academic language and to use vocabulary to engage in and express content learning.
17. The teacher consults with other educators to make academic language accessible to learners with different linguistic backgrounds.
18. The teacher engages learners in developing literacy and communication skills that support learning in the content area(s). S/he helps them recognize the disciplinary expectations for reading different types of text and for writing in specific contexts for targeted purposes and/or audiences and provides practice in both.
19. The teacher provides opportunities for learners to demonstrate their understanding in unique ways, such as model making, visual illustration and metaphor.

20. The teacher uses, designs or adapts a variety of classroom formative assessments, matching the method with the type of learning objective.

21. The teacher uses data from multiple types of assessments to draw conclusions about learner progress toward learning objectives that lead to standards and uses this analysis to guide instruction to meet learner needs. S/he keeps digital and/or other records to support his/her analysis and reporting of learner progress.

22. The teacher participates in collegial conversations to improve individual and collective instructional practice based on formative and summative assessment data.

23. The teacher engages each learner in examining samples of quality work on the type of assignment being given. S/he provides learners with criteria for the assignment to guide performance. Using these criteria, s/he points outs strengths in performance and offers concrete suggestions for how to improve their work. S/he structures reflection prompts to assist each learner in examining his/her work and making improvements.

24. The teacher matches learning goals with classroom assessment methods and gives learners multiple practice assessments to promote growth.

25. The teacher uses the provided curriculum materials and content standards to identify measurable learning objectives based on target knowledge and skills.

26. The teacher plans and sequences common learning experiences and performance tasks linked to the learning objectives, and makes content relevant to learners.

27. The teacher plans instruction using formative and summative data from digital and/or other records of prior performance together with what s/he knows about learners, including developmental levels, prior learning, and interests.

28. The teacher uses data from formative assessments to identify adjustments in planning.

29. The teacher identifies learners with similar strengths and/or needs and groups them for additional supports.

30. The teacher uses learner performance data and his/her knowledge of learners to identify learners who need significant intervention to support or advance learning. S/he seeks assistance from colleagues and specialists to identify resources and refine plans to meet learner needs.

31. The teacher uses data on learner performance over time to inform planning, making adjustments for recurring learning needs.

32. The teacher makes the learning objective(s) explicit and understandable to learners, providing a variety of graphic organizers, models, and representations for their learning.

33. The teacher analyzes individual learner needs (e.g., language, thinking, processing) as well as patterns across groups of learners and uses instructional strategies to respond to those needs.

34. The teacher poses questions that elicit learner thinking about information and concepts in the content areas as well as learner application of critical thinking skills such as inference making, comparing, and contrasting.

35. The teacher develops learners' abilities to participate in respectful, constructive discussions of content in small and whole group settings. S/he establishes norms that include thoughtful listening, building on one another's ideas, and questioning for clarification.

36. The teacher observes and reflects upon learners' responses to instruction to identify areas and set goals for improved practice.

37. The teacher gathers, synthesizes and analyzes a variety of data from sources inside and outside of the school to adapt instructional practices and other professional behaviors to better meet learners' needs.

38. The teacher uses technology and other forms of communication to develop collaborative relationships with learners, families, colleagues and the local community.

Notes

[1] PPAT responses are independently scored by two raters and the ratings are averaged. Therefore, in cases not involving adjudication, step-level scores of 1.5, 2.5, and 3.5 are possible.

[2] The reported PPAT score is (Task 2) + (Task 3) + (2 * Task 4).

[3] The recommended passing score shared with the panel was 40.5, the panel's mean rounded to the next highest available score. Following the full analysis of the study data, the median, 40.0, was selected as the recommended passing score.

[4] The SEJ for the median of the panelists' judgments is approximately 25% larger than the SEJ for the mean given the susceptibility to sampling fluctuation (MacCann & Stanley, 2004).