Listening. Learning. Leading.®

R & D Connections

No. 17 • November 2011

Setting Standards on *The Praxis Series*™ Tests:

A Multistate Approach

By Richard J. Tannenbaum

Key Terms or Concepts

- Licensure — A process by which state agencies or other governing bodies grant individuals the legal permission to practice an occupation. A license signifies that the holder has demonstrated sufficient levels of knowledge, skills, and/or abilities to competently perform important occupational tasks.
- *The Praxis Series*™ tests — *Praxis I*® tests measure basic academic skills. The scores are most often used to inform decisions about entrance into teacher preparation programs. *Praxis II*® tests measure subject-specific knowledge, as well as general and subject-specific teaching skills. The tests are taken by individuals entering the teaching profession as part of the licensure process required by many states.

continued on p. 2

When people speak of a "passing score" on a licensure test, they are talking about a *cut score* — the point on the score scale that separates those who qualify for licensure from those who do not.

The process used to recommend cut scores is known as *standard setting*. It usually involves a panel of experts reviewing the test content, defining the minimal knowledge and/or skills needed to qualify for licensure, and identifying the test score necessary to meet the qualification requirement.

Traditionally, standard setting for *The Praxis Series*™ licensure tests has been done state by state. A state department of education assembles a panel of educators who meet once to recommend a passing score for a specific test. ETS staff facilitates the standard setting. The state department of education then presents the recommendation to its state board, which sets the operational passing score. State boards meet only a few times each year to consider and respond to a range of issues and concerns. Departments of education therefore prefer to approach their respective boards with passing-score recommendations for several tests at a time. Three issues emerge from the traditional state-specific, one-panel-per-test standard-setting approach:

- A state must recruit a sufficiently large number of representative educators for each test-specific panel. Raymond & Reid (2001) suggest between 10 and 15 educators, and Zieky, Perie, & Livingston (2008) caution that a panel with fewer than eight members may not be defensible. For some licensure areas (e.g., mathematics, social studies, English Language Arts), these numbers may not present an issue for a state, but for other areas (e.g., economics, business education, physics), assembling even eight educators may not be feasible.
- Having one panel of educators make the passing-score recommendation leaves open the question of whether other panels of educators would have recommended a comparable passing score — i.e., the issue of reliability. Because passing a licensure test is one of the requirements for eligibility to teach in a state, evidence supporting the reliability of the passing score is important. There is no direct measure of reliability when — as is common practice — only one panel is assembled to make the passing-score recommendation. The reliability will then have to be approximated using the standard error of judgment.

Editor's note: *Richard J. Tannenbaum is Director of the Education and Credentialing Research Center in ETS's Research & Development division.*

Key Terms or Concepts

continued from p. 1

- Passing Score — The minimum score needed to pass a test. A passing score is also referred to as a *cut score* or *qualifying score*.
- Standard Setting — A variety of systematic, judgment-based processes that identify minimum test scores that separate one level of performance from another.
- Modified Angoff — A method of standard setting where panelists judge the probability that a minimally qualified test taker would answer a multiple-choice question correctly. Alternatively, panelists may judge the percentage of minimally qualified test takers who would answer the multiple-choice question correctly.
- Extended Angoff — A method of standard setting in which panelists judge the score that a minimally qualified test taker would earn on a constructed-response question. In a variation known as the Mean Estimation Method, panelists may judge the average score that would be earned by minimally qualified test takers.

- Passing scores for the same *Praxis*™ test tend to vary — sometimes widely — among states. A passing score reflects the values and expectations of those recommending the score and of those (on the state board) setting the operational score. The fact that passing scores often vary across states restricts the opportunity for teacher mobility across states, which, as Darling-Hammond (2010) suggests, undermines the equitable distribution of teacher quality.

This paper describes a multistate standard-setting approach designed to address the issues presented by the traditional approach for recommending passing scores on *The Praxis Series* licensure tests.

Multistate Standard Setting

Two design features distinguish the multistate standard-setting approach from the more traditional state-by-state approach for setting standards on *Praxis* tests.

The first distinguishing feature is that educators representing several states jointly participate in recommending the passing score for the test being considered. This has two benefits. One, it reduces each state's recruitment burden: Rather than having to assemble 10 to 15 educators on a panel, each state may only need to contribute up to four educators in the multistate approach. Two, the passing-score recommendation will reflect a more diversified perspective.

The second distinguishing feature is that two panels are formed from the same group of states for each test, and each panel makes a separate passing-score recommendation. The two panels permit a direct determination of the reliability of the passing-score recommendation, which is unique for teacher licensure. It is more common for a single state to bring together only one panel of educators to recommend a passing score for that state. This is likely due to the difficulty of recruiting a sufficient number of educators for more than one panel. But, as noted above, the multistate approach reduces any one state's recruitment burden; hence we are able to assemble two panels to recommend a passing score on the same test.

However, the multistate process does not use any new method of standard setting. The core methodologies — a modified Angoff for multiple-choice items and an extended Angoff for constructed-response items — are well established and widely used in setting standards on teacher licensure tests. This paper does not describe these methods in detail as several of the sources in the literature (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Tannenbaum & Katz, in press; Zieky et al., 2008) are available to readers who wish to study them further.

These standard-setting methods have long been in use for *The Praxis Series* tests. State departments of education and their state boards — responsible for setting the operational passing scores — are familiar with and accept these methods. It was an explicit goal to maintain these same standard-setting methods in developing the multistate standard-setting approach while still enabling us to:

- reduce the burden on any one state for recruiting educators to serve on standard-setting panels;

"The fact that passing scores often vary across states restricts the opportunity for teacher mobility across states, which, as Darling-Hammond (2010) suggests, undermines the equitable distribution of teacher quality."

- determine the reliability of a passing-score recommendation by having two panels of educators; and
- support a more uniform, national perspective regarding passing scores for initial teacher licensure.

Prerequisite Policy Groundwork

Setting a standard is, in effect, setting a policy (Kane, 2001), in this case a policy about the testing requirements for teacher licensure — the type and amount of knowledge and skills that beginning teachers need to have. Cizek & Bunch (2007) aptly note, "All standard setting is unavoidably situated in a complex constellation of political, social, economic, and historical contexts" (p.12). As we began to think about the multistate standard-setting process, this "truism" reinforced the need for us to present the concept to the state departments of education *before* any attempt to implement it. We needed to make sure that the state departments of education understood what we were proposing and why. We also wanted them to have the opportunity to raise issues or ask questions. In this regard, we implemented two strategies to secure state acceptance.

First, we conducted a series of webinars explaining how we envisioned the multistate process. Approximately 20 state departments of education participated. The webinars were instrumental in allaying states' concerns about changing a process that was familiar to them.

Second, we invited them to observe the first two multistate studies we had conducted for the School Leaders Licensure Assessment (SLLA) launched by ETS in September 2009. Several state directors or their designees attended these two studies to see firsthand how the studies were conducted and the nature of the interactions among educators from the states represented on the panels. State observers regularly attended multistate studies throughout the first year of implementation.

Overview of the Multistate Process

There are certain elements of panel-based standard setting that contribute to the quality and reasonableness of passing-score recommendations (Tannenbaum & Katz, in press). Several of these are part of the multistate approach and include having the panelists:

- take the test to become familiar with its content;
- construct a performance level description — the minimal knowledge and skills expected of a candidate who is qualified to be licensed;
- receive training in the standard-setting method(s) and the opportunity to practice making judgments; and
- engage in two rounds of standard-setting judgments.[1]

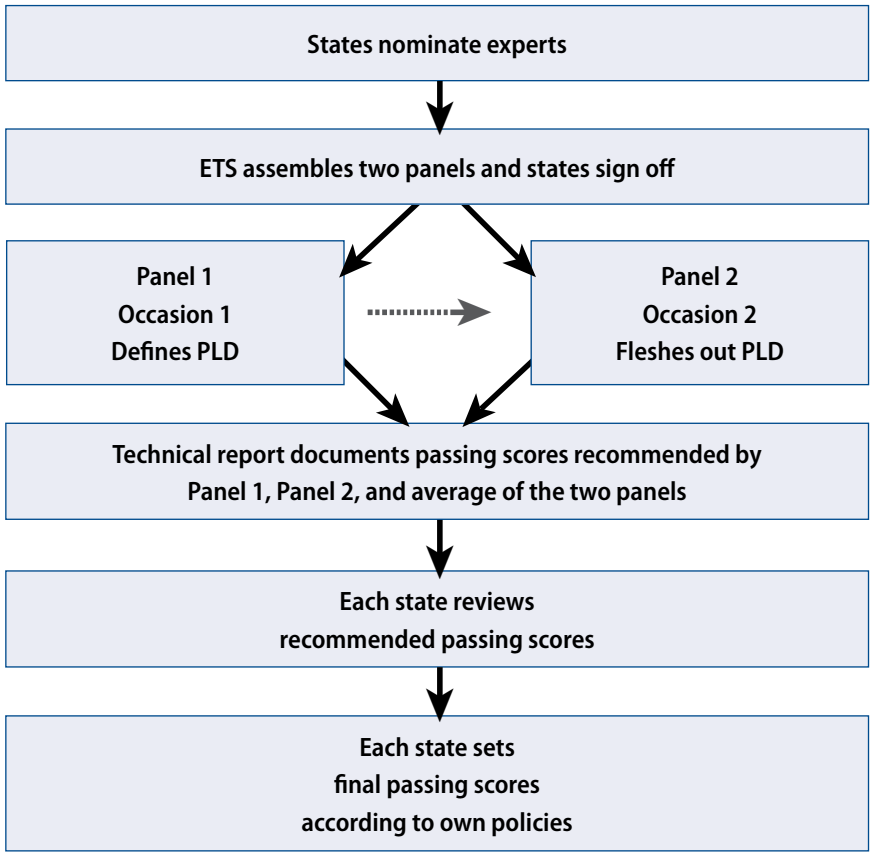
[1] Standard setting for *Praxis* tests occurs before test administration, as states tend to prefer to have passing scores in place prior to administration. This means that item- and test-level data are not available to inform the passing-score recommendation. The feedback between rounds is based solely on the educators' standard-setting judgments.

"The multistate process enables us to complete the standard setting for all states in the same brief time period, expediting the first step in each state's adoption cycle — the presentation of a recommended passing score to its board."

Each state interested in adopting the test for licensure nominates educators to represent the state. We encourage each state to nominate up to six educators — four teachers and two teacher education faculty members. They should come from different settings and be diverse with respect to gender, race, and ethnicity. Two panels are formed from the cross-state pool of nominees so that the composition and representation of the panels are comparable. We then contact each state to review the educators selected from that state, and ask each state to either approve the selection or suggest alternative educators. For the multistate process, each panel includes up to 25 educators. This number is greater than the range suggested by Raymond & Reid (2001) to bolster state representation, but it still supports interaction among the educators during the standard-setting process. This means that up to 50 educators may contribute to the passing-score recommendation for a test.

The two panels meet on different occasions, often within the same week. This expedites the test-adoption process for states. In the more traditional state-by-state model, states are necessarily placed in a queue for standard setting (due to resource limitations both on our part, and that of the states), which extends the timeframe for states to adopt *The Praxis Series* tests. The multistate process enables us to complete the standard setting for all states in the same brief time period, expediting the first step in each state's adoption cycle — the presentation of a recommended passing score to its board.

Figure 1: Multistate Standard-Setting Process



"The value of having a second panel of educators is that the number of educators contributing to the passing-score recommendation increases and we obtain a direct estimate of the reliability of the passing score."

The Performance Level Description

Because it meets first, Panel 1 has primary responsibility for constructing the performance level description, which delineates the minimal knowledge and skills expected of a candidate to pass the test. The objective of the standard-setting task is for the panel to identify the test score likely to be earned by a candidate who just meets the expectation expressed by the description.

The process of developing a performance level description roughly works like this: After having taken the test and discussed its content, the educators on Panel 1 are formed into two or three subgroups, and each of them independently constructs a performance level description. The entire panel then reviews and discusses the individual performance level descriptions in order to reach a consensus on a final performance level description. In general, we devote two hours to this work. The final description is printed so that each educator has a copy. The panel then completes the standard-setting task. Panelists respond to evaluation surveys after training and practice, as well as at the conclusion of the standard-setting session. The survey responses address the quality of the implementation and the reasonableness of the panel's recommended passing score.

Panel 2 can begin its work when this is done.

The value of having a second panel is that the number of educators contributing to the passing-score recommendation increases and that we can obtain a direct estimate of the reliability of the passing score. In this instance, reliability addresses the question of how close the recommended passing score of a second panel of educators would be to that of the first panel, if both followed the same standard-setting procedures and used the same performance level description. The key here is that our design maintains the consistency of the performance level description (the performance expectation) between the panels. The standard-setting method and the performance expectation remain constant, with only the particular educators on the two panels varying. The educators from Panel 2 take the test and discuss its content, just as the educators on Panel 1 did. However, the educators on Panel 2 receive the performance level description from the first panel rather than having to construct a new description. They are informed of the reason for this — to maintain consistency of the performance expectation with the first panel — and told what they need to do in order to "internalize" the meaning of the description. The work then proceeds in three steps:

Step one: The educators on Panel 2 discuss the performance level description as a group. The same researcher who facilitated Panel 1 also facilitates Panel 2, so that the "history" of Panel 1's performance level description can be shared as needed.

Step two: The educators are formed into subgroups, and each subgroup is asked to develop *critical indicators* for each of the knowledge or skill statements in the performance level description. A critical indicator is a signal to the educators that an individual is likely to have the defined knowledge or skill. The indicators are intended to help clarify the meaning of the knowledge or skill statement — to "flesh it out." Each subgroup is asked to generate two or three indicators for each knowledge or skill statement. The indicators are then presented to the whole panel for discussion, and a final set of indicators for each performance level description statement is documented.

"Over 90% of the panelists strongly agreed *that they understood the purpose of the study and that the standard-setting training they received was adequate. Nearly three-quarters (73%) of the panelists* strongly agreed *that the standard-setting process was easy to follow."*

Step three: The indicators are not intended to be exhaustive, but to illustrate what the statement means. Hence, only a few indicators for each statement are necessary. Table 1 presents examples of indicators associated with one knowledge or skill statement each from a performance level description for a *Praxis* Physical Education licensure test and for a general pedagogy licensure test (Principles of Learning and Teaching: Grades 7–12). The panel reviews the indicators for internal consistency before finalizing them — verifying that the entire set of indicators are related to the knowledge or skill statement, and that it has not changed the fundamental meaning of the statement. The panel then completes the same standard-setting task and evaluation surveys that Panel 1 completed.

Table 1. Examples of Critical Indicators

Test	Performance Level Description Statement	Indicators
Physical Education	understands individual and group motivation and behavior to foster positive social interaction, active engagement in learning, and self-motivation	classroom rules demonstrate positive social interaction models positive social interaction uses "do nows" (instant activity cards) that promote positive engagement
PLT (7–12)	understands how learner variables and areas of exceptionality affect student learning	identifies a variety of learning styles in each classroom modifies instruction and communication methods to meet a recognized need

Documentation

Each participating state department of education receives a technical report that documents the characteristics and experiences of the educators on the two panels, the methods and procedures each panel followed in arriving at the passing-score recommendations, and the round-by-round results for each panel. Once a state has received the report, it goes through its particular process to determine the final passing score to be set.

Quality Metrics

One indicator of the quality of the multistate standard-setting process comes from the panelists' responses to the final evaluation. Panelist evaluations are a credible indicator of the validity of the standard-setting implementation (Cizek, Bunch, & Koons, 2004; Kane, 1994). Table 2 presents results from surveys of more than 530 panelists across 16 tests. The tests cover art, business education, English, general pedagogy, physical education, school leadership, special education, technology education, teaching reading, and world languages. Table 2 summarizes the results for questions having to deal with panelists' understanding of the purpose of the standard-setting study, the adequacy of

the standard-setting training, and the ease of completing the standard-setting task. The responses were on a scale ranging from *strongly agree* to *strongly disagree*. The panelists were also asked to indicate whether they believed the recommended passing score was *about right, too low,* or *too high*.

Table 2. Responses to Final Evaluation

	% Strongly Agree	% Agree	% Disagree/ Strongly Disagree
I understood the purpose of the standard-setting study.	92.44	7.56	0
The training was adequate for me to complete the standard-setting task.	88.75	11.07	.18
The standard-setting process was easy to follow.	73.01	25.69	1.29
	% About Right	**% Too Low**	**% Too High**
How reasonable was the recommended passing score?	88.39	9.74	1.87

Over 90% of the panelists *strongly agreed* that they understood the purpose of the study, and believed that the standard-setting training they received was adequate. Nearly three-quarters (73%) of the panelists *strongly agreed* that the standard-setting process was easy to follow. We had expected a lower percentage here as standard setting is a novel activity for most educators; nonetheless, the positive result attests to the perceived quality of the standard-setting implementation. The percentages for these three questions approach 100 if the responses *strongly agree* and *agree* are combined. Close to 90% of the panelists indicated that the recommended passing score was *about right* and close to 10% indicated that it was *too low.*

A second indicator of quality is the reliability of the recommended passing score. The use of two panels is an explicit feature of the multistate standard-setting process. This increases the number of educators contributing to the passing-score recommendation, leading to a more stable recommendation, but it also permits a direct estimate of the reliability of the passing-score recommendation.

The recommended passing score is the Round 2 mean for a panel, so two means are available for each test (one for Panel 1 and one for Panel 2). Brennan (2002) provides a way to calculate a standard error of a mean when there are two observations, as is the case in the multistate approach. The standard error is the absolute difference between the two means (recommended passing scores) divided by two. Sireci, Hauger, Wells, Shea, & Zenisky (2009) suggest that a value of less than 2.5 indicates that other panels of educators would recommend comparable passing scores. The standard error is less

than 2.5 in all 16 instances. The lowest value was 0.43 and the highest was 2.14; the average value was 1.29. This indicates that the recommended passing scores should not vary significantly across other panels of educators.

A third indicator of quality in a multistate process is the variability in passing scores across states. A reduction in variance indicates a higher potential for teacher mobility, and preliminary evidence points in this direction. Between 2008 and 2010, 13 states that participated in single-state studies set passing scores for 37 *Praxis* tests. The average percentage of change from the panel-recommended passing score was approximately 5 scaled points. Between 2009 and 2010, 26 states that participated in multistate studies set passing scores for 10 *Praxis* tests.[2] The average percentage of change from the panel-recommended passing score in these instances was approximately 1 scaled point.

Conclusion

Traditional standard setting for *Praxis* teacher licensure tests is done on a state-by-state basis, with each state assembling one panel of educators on one occasion to recommend a passing score. This places a burden on each state to recruit a sufficient number of educators to serve on a panel, leaves open to question whether other panels of experts would recommend a similar passing score, and often leads to variation in passing scores across states. This article outlines a multistate standard-setting process that addresses these issues.

References

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.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*, 31–50.

Darling-Hammond, L. (2010, October). *Evaluating Teacher Effectiveness: How Teacher Performance Assessments Can Measure and Improve Teaching*. Retrieved May 10, 2011, from Center for American Progress (http://www.americanprogress.org).

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.

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

[2] Not all states that participated in standard setting set final passing scores in this time period.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Raymond, M. R., & Reid, J. R. (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: Lawrence Erlbaum.

Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C., & Zenisky, A. L. (2009). Evaluation of the standard setting on the 2005 Grade 12 National Assessment of Educational Progress mathematics test. *Applied Measurement in Education*, *22*, 339–358.

Tannenbaum, R. J., & Katz, I. R. (in press). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*. 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: ETS.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001
email: **RDWeb@ets.org**

Editors: Jeff Johnson and Hans Sandberg
Copy Editor: Eileen Kerrigan
Layout Design: Sally Acquaviva

Visit ETS Research & Development on the web at **www.ets.org/research**

Follow ETS Research on Twitter **(@ETSresearch)**