



National Council on
Teacher Quality

National Council on Teacher Quality
1032 15th Street NW #242
Washington, DC 20005

202-393-0020
www.nctq.org

October 30, 2024

Commission on Teacher Credentialing
651 Bannon Street, Suite 600
Sacramento, CA 95811

To the California Workgroup to Review the Design and Implementation of Teaching Performance Assessments:

Every student deserves to have effective teachers in every year of their school career, and that right should not be compromised when they have a first-year teacher. Ensuring the effectiveness of new teachers is especially important to address educational equity, because students of color and students living in poverty are most likely to have novice teachers.

Licensure tests, which apply comprehensive and standardized expectations across all incoming teachers, set guardrails so that even students with first-year teachers have access to a high-quality education. These tests also clearly indicate to prep programs what they are expected to teach their candidates, and they provide a source of data to verify that programs are providing that instruction effectively.

The National Council on Teacher Quality (NCTQ), a national research and policy nonprofit, is pleased to see that the California Commission on Teacher Credentialing's (CTC) workgroup is examining Teaching Performance Assessment (TPA) requirements. We would like to share some considerations for these assessments based on our review of research and lessons learned from other states.

Clearly define what the assessment can and cannot effectively evaluate

A performance assessment can serve a valuable purpose, offering insight into an aspiring teacher's ability to effectively deliver instruction in a real classroom. However, performance assessments, like any assessment, have limitations. It is important that this workgroup clearly defines and communicates what the TPA intends to measure as well as what is outside the scope of this assessment.

For example, because the TPA design typically allows candidates to identify specific content they want to teach, and because candidates must only provide one or a few samples of instruction on select topics, TPAs generally do not offer a valid or reliable measure of candidates' content knowledge in a topic. Whereas content tests are typically designed to ask a large number of questions that sample candidates' knowledge across the breadth of a subject area, the TPA design does not assess the same breadth of content. **For that reason, the TPA should not replace**



essential content exams like the RICA, which is considered a strong measure of aspiring teachers' knowledge of reading.

Similarly, while the TPA may provide insight into candidates' pedagogy, it may or may not provide information about a candidate's ability to manage a classroom and promote a positive environment. Many TPA assessments allow candidates to submit videos of themselves leading a small group in instruction, rather than providing whole-class instruction or facilitating learning across the entire class. This approach will not provide information about whether a candidate can effectively manage an entire class of students.

Consider also whether and how the TPA can provide insight into a candidate's readiness to be an effective first-year teacher, based on alignment with the state's (or districts') evaluation rubric for in-service teachers. Massachusetts took this approach with their Candidate Assessment of Performance. With their assessment, they evaluate candidates on a subset of elements used to evaluate in-service classroom teachers to determine whether student teachers are "ready to teach."

Investigate and limit burden on teacher candidates

Recent research has documented the high burden that TPAs present to teacher candidates in terms of both time and cost, which may reduce the number of people who ultimately earn teaching licenses. The burden may also fall disproportionately on candidates of color, a factor the workgroup should well consider. While any rigorous assessment will, by necessity, require that candidates take time to prepare, the workgroup can identify ways to mitigate the burden on candidates. These steps can include:

- Taking stock of all required elements of the TPA and estimating the time required by each,
- Clearly defining the purpose and desired outcome of each required element and revisiting whether each required element is essential, and
- Identifying other ways to vet essential knowledge (e.g., separately testing content knowledge through content licensure tests, shifting some observation requirements into clinical practice requirements rather than part of the TPA).

The workgroup should also consider ways to mitigate the cost of the assessment for candidates, including offering vouchers for initial or second attempts or allowing prep programs to cover the cost through student fees, which may allow candidates' financial aid to cover the cost of the assessment.

Research by Chung and Zou also found that prep programs' instruction was not always aligned with the expectations of TPAs, creating additional confusion and work for aspiring teachers. Ensuring this coherence in expectations from the outset (by both encouraging prep programs to



infuse TPA standards throughout prep program coursework and by designing TPA standards to align with expectations from prep programs as well as the state's standards for teacher candidates) can help candidates more efficiently prepare to meet the expectations of the TPA.

Ensure reliability in scoring

For a performance assessment to indicate whether an aspiring teacher is prepared to earn a license and become a teacher of record, it must be both valid (scores represent a candidate's true ability) and reliable (scores and scoring processes are consistent across all raters, all candidates, and over time).

Ensuring that scores on a performance assessment are valid and reliable is an inherently challenging task. A pointed study by Drew Gitomer and colleagues called into question the reliability and validity of scoring on the prevalent edTPA. While the developer of the edTPA disputed some of the critiques, this analysis does raise important considerations for the development and scoring processes for any performance assessment.

In addition to conducting robust reliability testing that meets prevalent psychometric standards, NCTQ recommends the following:

- Select raters from an array of different preparation programs (i.e., do not allow prep program faculty to evaluate their own candidates) and from outside of prep programs (e.g., highly effective teachers, school leaders, or other stakeholders). This variety helps reduce the chance that raters are biased in favor of their own candidates or of teacher candidates more broadly.
- Provide raters with training to calibrate scores. Raters should score multiple sample assessments, and they should not be qualified to score actual assessments until their scores consistently match the sample score. Raters should also go through a process to re-certify that they are still calibrated on scoring; this should be done on a specified timeline and be required at least every two years. Note that a past study comparing different TPAs allowed in California found some limited opportunities to strengthen the training provided to raters.
- As raised by Gitomer et al., assign different raters to score different portions of each candidate's assessment, so that no single rater is scoring a candidate's entire assessment. This process can help reduce the likelihood that a candidate's score is deeply affected by bias or inconsistency from one rater. Gitomer notes that this was a standard practice of another highly regarded assessment for National Board Certification: The original NBPTS assessments included 10 separate tasks, each scored by two raters.
- Establish a protocol for the state to review scores throughout the scoring process. This protocol should include looking for drift over time (e.g., scores are drifting higher or lower); differences in scores between raters (e.g., one rater gives consistently lower than average



scores); differences in scores based on the race, ethnicity, language, or gender of the candidate or based on their class's composition; and differences by category of rater (e.g., prep program faculty, school leader), among others that the TPA developers deem relevant.

- As another means to check the validity and reliability of scoring, randomly sample a portion of assessments to be scored by two different raters. All raters should conduct at least some of this “double-coding” and be paired with a rotating array of fellow raters. This process allows for comparisons among all raters and enables further checks that the raters are applying consistent approaches to scoring.
- Finally, conduct a rigorous check for bias throughout the development and roll-out of the TPA. Processes should include having a diverse group of stakeholders review individual assessment items for bias in how they are phrased or constructed; and reviewing both pilot and full-scale data for evidence of bias or inconsistencies in scoring based on the race, ethnicity, gender, or other characteristics of the candidates, raters, or classroom students. An analysis of past TPAs finds some differences in mean scores, although not pass rates, by race.

Build in data collection and data sharing

If performance assessments align with the standards that newly licensed teachers should meet and with the instruction teacher prep programs provide, the data from these assessments can offer rich insight into incoming cohorts of teachers. For example, prep programs could use the data to identify areas in which their candidates persistently struggle and then modify coursework to better target those areas. The state could identify strengths and limitations of incoming cohorts of teachers and provide additional support (e.g., through targeted mentoring) as well as clarify standards and expectations for prep programs.

For stakeholders to use the data in this way, a consideration when developing the TPA must be how to build a data management system that is easy for stakeholders to use. This system should allow prep programs to view:

- Data on any individual candidate in their program, or information on groups or cohorts of candidates in their programs (e.g., to allow comparisons across years, across candidates enrolled in different sections of the same course, or across demographic groups within the program).
- Detailed data (e.g., overall scores, component scores, and subcomponent or item-level scores).
- Comparisons between their prep programs and other similar programs (e.g., to compare outcomes across all graduate secondary mathematics programs in the state).



National Council on
Teacher Quality

National Council on Teacher Quality
1032 15th Street NW #242
Washington, DC 20005

202-393-0020
www.nctq.org

This system should also allow state-level stakeholders (e.g., members of the California Commission on Teacher Credentialing) to see data across the state, including:

- Overall trends in scores (e.g., the percentage of candidates scoring at each level, as well as that data disaggregated by race/ethnicity, gender, certification area, and other topics of interest).
- Score trends broken out by institution or teacher prep program, both aggregated and disaggregated.
- Aggregated and disaggregated data on component- and subcomponent-level scores.

Measure the predictive validity of the assessment

Typically, the ultimate goal of an assessment required for teacher licensure is to distinguish between who is more likely to be an effective classroom teacher and who is less likely. During the pilot test and early years of the requirement for the TPA (and periodically in the years after), the state should examine whether and how well the assessment predicts future teachers' effectiveness. Researchers have conducted similar evaluations in other states. Understanding the predictive validity of the assessment can help efforts to refine the test in the future so that it better achieves its goals, and can provide support to either maintain or revisit this licensure requirement. Including plans for this future analysis at the outset can ensure that the state, prep programs, and test developers are collecting the relevant information in a way that can be used for this analysis.

In conclusion, NCTQ believes that TPAs hold great promise for improving the quality and readiness of new teachers entering the classroom, but that promise depends upon the design of the assessment. A well-designed TPA should be clear in purpose, scored through a process that rigorously vets reliability and validity, reviewed for bias, and able to provide data that can drive improvement in preparation and can hone future iterations of the test. These steps can help ensure that any burden candidates take on is limited but—more importantly—is worthwhile.

Thank you for considering this input, and please do not hesitate to reach out with any questions.

Best,

Heather G. Peske, Ed.D.
NCTQ president