
4A

Information

Professional Services Committee

Increasing the Reliability (Scoring Consistency) of Candidate Results on the Teaching Performance Assessment (TPA)

Executive Summary: This agenda item presents considerations for the potential verification and improvement of the reliability (scoring consistency) of candidate results on the Teaching Performance Assessment.

Policy Question: This agenda item raises several issues for Commission review and direction. How does the Commission wish to proceed to address those issues?

Recommended Action: For information only

Presenters: Phyllis Jacobson and Mike Taylor, Consultants, Professional Services Division

Strategic Plan Goal: 1

Promote educational excellence through the preparation and certification of professional educators

- ◆ Sustain high quality standards for the preparation and performance of professional educators and for the accreditation of credential programs

August 2012

Increasing the Reliability (Scoring Consistency) of Candidate Results on the Teaching Performance Assessment (TPA)

Introduction

This agenda item continues the discussion of issues relating to the implementation of the Teaching Performance Assessment (TPA) requirement begun at the Commission's April 2012 meeting (<http://www.ctc.ca.gov/commission/agendas/2012-04/2012-04-6B.pdf>). This agenda item addresses the issue of increasing the reliability (i.e., scoring consistency) of candidate results on the assessment. An Executive Summary of the options presented in this agenda item is provided in Appendix D.

Background

As of July 2008, California statute (Education Code §44320.2) requires all candidates for a Preliminary Multiple and Single Subject Teaching Credential to pass an assessment of their teaching performance with K-12 public school students as part of the requirements for earning a preliminary teaching credential. Between 2003 and 2008, several teaching performance assessment models had been developed and were being implemented on a voluntary basis by individual teacher preparation programs.

Approximately 23,065 candidates took the TPA in 2010-2011. Of these, 63% took the CalTPA; 33.6% took the Performance Assessment for California Teacher (PACT); and the remaining 3.4% took the Fresno Assessment of Student Teachers (FAST).

Statutory Responsibility for Assuring TPA Scoring Reliability

Under the Education Code, the Commission has several responsibilities with respect to data collection and analysis relative to TPA results. Section 44320.2 requires the following with respect to analysis of the reliability of assessment scoring and the analysis of candidate score results:

44320.2 (d) Subject to the availability of funds in the annual Budget Act, the commission shall perform all of the following duties with respect to the performance assessment:

(4) Initially and periodically analyze the validity of assessment content and the reliability of assessment scores that are established pursuant to this section.

(7) Collect and analyze background information provided by candidates who participate in the performance assessment, and report and interpret the individual and aggregated results of the assessment.

Part I: Current Process for Looking at TPA Score Reliability

Overview of the Current Score Reliability (Scoring Consistency) Process

Following statewide mandatory implementation of the teaching performance assessment requirement in 2008, procedures were put into place to review the reliability of candidate scoring

outcomes across programs and models. The process recommended by the TPA Implementation Task Force, which at the time was the guiding body for the statewide TPA implementation and included statisticians from each of the three approved TPA models, was to require programs to double-score a minimum of 15% of candidate TPA responses and to look at the rate of scorer agreement for the double-scored responses within the program and across program assessors.

Currently, the rescore process is managed by the preparation programs, and the resulting data are looked at by the local preparation programs. The data and the program's analysis of the impact of the data are reported to the Commission through the accreditation process, specifically as one of the data points required in the Biennial Report.

Critical Role of Scorer Training, Calibration, and Recalibration in the Score Reliability Process

Critical to the reliability of the scoring process is the training and calibration of scorers. Following successful initial calibration, scorers are required to recalibrate on at least a yearly basis (if they have not been consistently scoring during the year). The calibration process consists of having scorers independently read and score actual candidate responses that have been previously scored by experts. The scorer must match the score awarded by experts at the level determined by the TPA model developer.

Each of the models has developed its own internal scorer training model, which specifies the qualifications for who can be trained as a scorer, the content of the training, who does the training, and what initial calibration standard the scorers have to meet in order to successfully complete the training and be allowed to score actual candidate TPA responses. Following initial calibration, scorers must maintain their calibration status in order to continue to score candidate TPA responses. Each approved TPA model conducts its recalibration process as determined by the model developer. All three models have developed at least one version of an online recalibration process which is maintained and/or overseen by staff of the respective model developers.

Additional Factors Affecting TPA Scoring Consistency (Reliability)

a. Local implementation of the TPA

TPA implementation takes place at the local teacher preparation program level. Program sponsors must implement the selected model as that model was designed and validated by the model's developer. With respect to scoring, programs are responsible for:

- identification and training of qualified scorers of candidate performance
- assuring that candidate performance is assessed by trained and calibrated scorers in a manner that is fair and reliable
- maintaining candidate, scorer, and outcomes data
- using TPA-related data for program improvement purposes

Thus, the TPA statewide system relies on the ability of each local teacher preparation program to carry out these responsibilities in a consistent manner and to select and use only those scorers who are highly qualified and who meet a high standard of continuous calibration. In practice, it is less clear that each preparation program consistently meets these standards. As initial training

gets further replicated with new lead trainers and new program personnel as well as new scorers over time, there is a risk and a likelihood of dilution of quality throughout the system, absent a statewide process for assuring the maintenance of the necessary level of quality over time and for assuring implementation of the model as designed by the developer.

In the event that a program identifies a scorer who is not maintaining sufficient calibration, there is the further complication of replacing that scorer in a timely manner with another trained and calibrated scorer, assuming that the program has sufficient staff resources to do so. In a smaller preparation program with limited staff, this factor can become a significant scoring quality and reliability issue.

Two of the three models use a Lead Trainer approach, whereby some scorers become qualified and/or authorized to train other scorers. The Performance Assessment for California Teacher (PACT) has implemented an online approach for experienced scorers to become trainers; CalTPA conducts Lead Assessor training in person for purposes of quality control across the model's participating programs. The Fresno Assessment of Student Teachers (FAST) conducts all of its own training for assessors. The Lead Trainer approach presents an additional possibility for dilution of quality and reliability of scoring, as the lead trainers become further distanced over time from the model developers, and the issue of maintaining the quality of the trainers themselves becomes an additional complication within the overall scoring system.

b. Scoring differences among the three TPA models affect comparisons of candidate outcomes

Although all three models measure candidate performance against the Teaching Performance Expectations (TPEs), there are some key differences in the scoring structure of each of the three models that affect candidate outcomes on the TPA:

- When the assessment is given, and thus scored, varies across programs. For example, the four tasks of the CalTPA are designed to be completed across the entire span of the credential program, whereas the PACT event typically takes place during the latter part of the program during student teaching. The timing of the assessment may influence the scoring of the assessment, since candidates with less program experience may not score as well as candidates who take the assessment towards the end of the program experience. This situation varies not only across models, but also within models at the level of individual program implementation.
- The method of deciding if an individual has passed the assessment varies across the models. The CalTPA and FAST models each utilize an overall passing score candidates need to meet. The CalTPA score is a single score for each of the four tasks. The four scores are reviewed within a compensatory model to determine if a candidate has met the overall passing standard. Passing status for the PACT models is based on decision rules based on performance on each of the 11 rubrics along with a minimum score. The determination of passing status varies not only across models, but also within models at the level of individual program implementation since programs may choose a higher standard than the minimum standard set by the model developer.
- Scoring reliability data can only be aggregated at the model level, so no overall “statewide” results can be provided. There is no common total candidate score that makes sense outside of the particular model each institution uses.

- At the end of any given academic year, the majority of enrolled candidates have not attempted to complete the full TPA. This factor makes it difficult to collect annual data on candidate and scorer outcomes.

The results of the different scoring conditions described above affect the candidate outcomes reported by individual programs, as shown in Table 1 below:

Table 1: Statewide TPA Passing Status by Demographic Variable

		Number who attempted all sections of the TPA by the end of 2009-10	Candidate attempted all sections of the TPA but did not pass one or more sections. No additional attempts are pending.		Candidate passed all sections of the TPA, one or more sections had to be repeated to pass		Candidate passed all sections of the TPA on the first attempt	
TPA Model	All Candidates	11,036	215	2%	1,515	14%	9,306	84%
	CalTPA	5,894	138	2%	1,222	21%	4,534	77%
	FAST	626	0	0%	84	13%	542	87%
	PACT	4,516	77	2%	209	5%	4,230	94%
Program Type	Traditional	8,557	99	1%	1,155	13%	7,303	85%
	Intern	1,248	98	8%	258	21%	892	71%
	Blended	441	3	1%	55	12%	383	87%
	Unknown	790	15	2%	47	6%	728	92%
Credent- tial Type	MS	5,530	130	2%	720	13%	4,680	85%
	SS	5,011	85	2%	716	14%	4,210	84%
	Dual	93	0	0%	24	26%	69	74%
	Unknown	4	0	0%	0	0%	4	100%
Gender	F	8,118	147	2%	1,007	12%	6,964	86%
	M	2,840	64	2%	500	18%	2,276	80%
	Unknown	77	4	5%	8	10%	65	84%
Ethnicity /Race	American Indian or Alaska Native	49	1	2%	10	20%	38	78%
	Asian	717	17	2%	83	12%	617	86%
	Black or African-American	282	11	4%	65	23%	206	73%
	Hispanic/Latino of any race	1,959	50	3%	325	17%	1,584	81%
	Native Hawaiian or Other Pacific Islander	68	0	0%	8	12%	60	88%
Ethnicity	White	5,392	90	2%	725	13%	4,577	85%

		Number who attempted all sections of the TPA by the end of 2009-10	Candidate attempted all sections of the TPA but did not pass one or more sections. No additional attempts are pending.			Candidate passed all sections of the TPA, one or more sections had to be repeated to pass		Candidate passed all sections of the TPA on the first attempt	
/Race <i>(continued)</i>	Two or more races	210	2	1%	32	15%	176	84%	
	Unknown	2,227	42	2%	242	11%	1,943	87%	
Native English Speaker	Yes	5,100	50	1%	670	13%	4,380	86%	
	No	733	13	2%	142	19%	578	79%	
	Unknown	5,158	144	3%	699	14%	4,315	84%	
Highest Degree Held	Associate	58	0	0%	9	16%	49	84%	
	Bachelor	7,906	158	2%	1,148	15%	6,600	83%	
	Master	429	11	3%	55	13%	363	85%	
	Doctorate	31	1	3%	4	13%	26	84%	
	Special, e.g. Juris Doctor	14	0	0%	0	0%	14	100%	
	None	226	3	1%	55	24%	168	74%	
	Unknown	2,371	42	2%	244	10%	2,085	88%	

As documented in Table 1 above, more than half of the candidates enrolled during the 2009-2010 academic year had not attempted all sections of the TPA by the end of that academic year. The percentages shown in Table 1 are the percent of candidates who attempted all sections of the TPA, not the percent of total enrolled candidates for each category.

As the table shows, most candidates who attempted all portions of the TPA passed on their first attempt (84%). However, it is not appropriate to directly compare the first time pass rates across all programs because of the differing conditions under which candidates may have taken the assessment. For example, in the CalTPA model, candidates take the different tasks at varying points in the program, starting from their early coursework, while in the PACT model candidates typically take the assessment later in the preparation sequence. Another factor affecting the overall passing rates is that some local programs permit a higher number of retakes for candidates than do other programs, where candidates may be limited to one additional attempt. Some programs may also counsel candidates out of the teacher preparation career choice early in the program, depending in part on TPA results, while other programs where the TPA occurs later in the preparation sequence may not counsel students out prior to completion of the TPA.

Additional observations regarding the data include:

- Candidates identified as being enrolled in a traditional preparation program pass the TPA the first time they attempt it at a higher rate than candidates identified as being enrolled in

an intern program. Candidates identified as being enrolled in a blended program have the highest first-time pass rate.

- Pass rates are very similar for both multiple-subject and single-subject candidates.
- Nearly three quarters of all candidates required to complete the TPA are female, and female candidates are passing TPA on their first attempt more often than male candidates.
- The fact that there are large numbers of candidates for whom no data was reported regarding ethnicity, native English speaker status, and highest degree held makes it difficult to draw any conclusions from the data in these fields.

c. Lack of a minimum acceptable rate of scorer agreement to verify reliability of TPA scoring

Currently, the Commission requires individual programs to rescore a minimum of 15% of candidate responses and to look at the rate of scorer agreement across that 15% sample. However, neither the Commission nor the individual models have set a minimum rate of scorer agreement that would verify the reliability of the TPA assessment scores and would establish a process that should ensue where reliability is low to improve the level of reliability. In addition, it is difficult to establish the reliability of assessment scores when there is only a small number of candidates assessed; for example, the 15% rescore requirement in a small program can result in fewer than 10 rescoring.

d. The relationship of the accreditation process to the TPA scoring consistency process

Accreditation is the Commission's avenue for assessing program implementation of its educator preparation standards. Since the TPA requirement is addressed within the Multiple and Single Subject program standards, the Commission's review of program implementation of the TPA currently occurs within the accreditation review process. Within that process, program documentation as well as onsite accreditation visits are intended to assure the Commission that programs are meeting the Commission's standards relating to the TPA by implementing the selected model in accordance with its design, including assuring the reliability of the assessment scoring by the program's trained scorers.

However, understanding the complexities of the three distinct TPA models as well as the psychometric principles relating to scorer training and scoring validity requires accreditation staff with appropriate background and experience in the TPA. Therefore, in order to provide expert review of information submitted by program sponsors relating to the implementation of the TPA within the ongoing accreditation process, a cadre of TPA experts has been identified to assist the work of the accreditation unit in reviewing documents relating to TPA implementation. Appendix A provides the Program Sponsor Alert (10-17) containing details regarding this process and the relationship of TPA to accreditation.

The Biennial Report process, however, may not be the most appropriate or effective method by which the Commission reviews the reliability of scoring across all TPA models and teacher preparation programs, for the following reasons:

- there is presently no standard data collection format for scorer data and for score agreement data for programs to use within the Biennial Report process.
- additional trained personnel would need to become part of the reading and evaluation of the TPA scorer calibration data reported by programs.

- the data relating to scoring consistency (reliability) is not provided in a format that would allow the data to be extracted from the Biennial Reports and aggregated across programs and TPA models in order to provide a statewide picture of the reliability of TPA scoring.
- since not all programs provide Biennial Reports each year, it is not possible to obtain a statewide picture on an annual basis through the Biennial Report process, or even to provide a consistent look over time across programs and models through this process.
- the Commission has little to no control as to the quality of the data, the completeness and timeliness of the data, and the accuracy of the data submitted in the Biennial Reports.

e. The Commission's role as both the state agency in charge of the TPA statewide process and a model developer responsible for one of the three approved models

The Commission has the statutory responsibility not only for overseeing the statewide implementation of the TPA and for reviewing and approving additional TPA models, but also for the ongoing development and implementation of one specific TPA model, the CalTPA. Some Commission staff work specifically on the CalTPA and represent the CalTPA on the various advisory bodies (i.e., the Users Advisory Committee and the CalTPA Steering Committee) while other staff focus more on the statewide implementation of the TPA across all models. It can be difficult to discern when the Commission is addressing the TPA in general and when the Commission is necessarily promoting the interests of the CalTPA as the developer/owner of this model. It can be difficult for the field as well to determine when the Commission is addressing all three models or the interests specifically of the CalTPA. This factor is an evitable result of the language of the statute which assigns the Commission both responsibilities. In essence, in regulating the statewide implementation of the TPA, the Commission is acting on behalf of the state as a whole but its actions also affect its own proprietary interest in and responsibilities for the CalTPA model.

f. The role and relationship between the Commission and the model developers

The Commission has the statutory responsibility for oversight of the TPA process, including reviewing and approving TPA models developed by other entities and for holding TPA models accountable for meeting the Assessment Design Standards (Appendix B). However, once the existing models have been approved by the Commission, a further or formal process for interacting with model developers over time has not been developed beyond the discussions held by the Users Advisory Committee.

g. The relationship between model developers and local teacher preparation programs implementing the TPA requirement

The three Commission-approved model developers have spent extensive resources to develop, validate, and assist local preparation programs in implementing the models. Commission standards require the local teacher preparation programs to "implement the model as designed." However, TPA models are not frozen in time; they evolve to meet the changing needs and conditions of teacher preparation programs. The Commission's process for helping models identify and meet challenges and issues arising from local program implementation is primarily through the efforts of the Users Advisory Committee, a body which has representation from all three models and model statisticians.

How much the model developers/owners can influence local implementation policy, and how much the model developer/owners can add requirements to programs for specific implementation activities, including such areas as scorer selection and training, for example, have not been clearly defined by the Commission or established through practice. The model developers/owners are further constrained by a lack of resources sufficient to modify, expand, or revise the existing TPA model in response to changing conditions such as, for example, changes in the CSTP and the TPEs on which the models were originally based. The more that the Commission adds requirements for model developers, the more costly for the model developers owners (including the Commission itself), and the less likely that these changes or requirements would be able to be implemented in the current fiscal climate.

h. Legal responsibility for and defensibility of TPA scores

Currently the TPA is implemented as a locally owned and operated process. Individual teacher preparation programs choose a TPA model to implement, identify and select their own TPA scorers, obtain or provide training for those scorers, prepare candidates for the assessment, schedule and administer the TPA, and score their candidates. Candidates not satisfied with the TPA process or the outcomes follow institutionally-established policies for addressing candidate concerns and/or complaints.

The more that the Commission (i.e., the state) becomes involved in regulating the TPA process, even with respect to increasing the consistency of scoring across programs and models, the greater becomes the potential responsibility and liability of the Commission for defending the scoring process and the resulting candidate outcomes. This issue addresses the complex balance between the Commission's regulatory role and legislative intent for a locally developed and administered assessment.

Part II: A Continuum of Potential Approaches to Increase the Scoring Consistency (Reliability) of TPA Candidate Outcomes

Introduction

The list below indicates potential approaches that could be taken along a continuum of actions towards assuring the reliability of TPA assessment scores, as required by statute. A further explanation of each of the bullet points follows the list.

- establishment of a minimum statewide rate of scorer agreement to verify scoring consistency
- increased statewide minimum number and/or percentage of rescoring
- increased statewide frequency of required scorer recalibration
- increased model developer supervision of scorer calibration/recalibration
- increased model developer supervision of lead trainer selections
- specified stratified random sample of rescoring
- require annual reporting in a format other than Biennial Reports (complete template)
- centralized rescoring (statewide or regional basis)
- centralized scoring for all models
- centralized administration and scoring for a single statewide model

Discussion of Potential Approaches Across a Continuum

Establishment of a minimally acceptable rate of scorer agreement to verify scoring consistency

Establishing a minimally acceptable rate of scorer agreement would provide a guideline for programs to identify if their scores are sufficiently reliable to support candidate outcomes that ultimately affect credentialing determinations. However, it is not clear that given the range of candidate enrollment across programs, there would be sufficient numbers of candidates in each program to obtain a statistically reliable sample of candidates and of scorers.

This issue aside, if the Commission were to want to establish a minimum rate of scorer agreement, there are three potential options:

- Individual preparation programs could establish their own minimally acceptable rate of scorer agreement.
- Each TPA model developer could establish a model-wide minimally acceptable rate of scorer agreement, and programs could use this rate or choose a higher rate. Since this would represent a change to the model requirements, the Commission would likely need to review and approve the model-established minimally acceptable rates of scorer agreement.
- The Commission could establish a statewide minimally acceptable rate of scorer agreement, and programs could use this rate or choose a higher rate.

A related consideration is what would be required of programs if they did not meet the minimally acceptable rate of scorer agreement. This issue is likely to result in higher costs for programs that do not meet minimally acceptable scorer agreement rates, since the programs would need to address the situation by taking actions such as, for example, retraining and recalibrating existing scorers; recruiting and training new scorers; contracting with other entities and/or scorers from other programs; reviewing the credentialing decisions made about candidates whose scores were found not to be in agreement, and other local approaches. Increased record-keeping for programs would also result from the need to track more closely the results of rescoring and of individual scorers whose work was rescored.

Increased minimum number and/or percentage of rescoring and of scorers included in the rescoring process, within each approved program

Currently the Commission requires 15% of the candidate TPA responses to be rescored (i.e., double-scored) in order to sample the reliability of the assessment score. Programs should look at the rate of scorer agreement based on this process in order to identify first, if candidate assessment scores are reliable, and second, if scorers are remaining calibrated.

However, to improve the data relating to scoring reliability and scorer reliability the Commission might choose to require either a minimum number of candidate responses to be rescored or a specific percentage of candidate responses to be rescored, depending on the size of the preparation program. For example, the current 15% rescore requirement for a small program of 45 candidates would result in approximately 7 rescoring, whereas the same requirement for a large program of 150 candidates would result in 23 rescoring. 7 rescoring is too small a number to make any significant conclusions regarding scoring reliability, and for a large program 23 rescoring may not be a sufficiently large sample to include the range of different scorers.

Along with the question of the appropriate number and/or percent of rescors is the question of an appropriate number of different scorers who should be included in the rescore process in order to view how the scorers across the program as a whole are performing.

Increased frequency of required scorer recalibration

Currently once scorers have met initial calibration requirements, they are required to recalibrate only once per year. It is not clear that this is sufficient to assure scorers are maintaining their calibration status over time, especially as not all scorers are included in the rescoring process. The Commission might want to require all models to increase the frequency of required scorer recalibration.

Increasing the frequency of required scorer recalibration, while it would help improve the reliability of assessment scores, would also have a cost in terms of (a) the model developers needing to more frequently update the online recalibration cases; (b) the programs needing to keep track of when each scorer needs to recalibrate and assure that the scorer successfully completes that process; and (c) additional scorer time and effort, which may also involve increased program costs for scorer services.

Increased model developer supervision of scorer calibration/recalibration

The quality of the scorers and of scorer training are both critical components to the eventual determination of score reliability. Currently, because the TPA is a locally owned and implemented process, individual preparation programs identify the qualifications for scorers and select scorers for training. There is currently limited to no input from model developers regarding the specific individuals whom the local programs select as scorers.

Model developers typically provide initial scorer training and also provide an online version of scorer recalibration. Some local programs, however, also have trainer of trainers who may provide initial scorer training and/or recalibration activities. It is also possible that model developers may not ever see and/or interact with some and/or all of the scorers, and thus it is difficult for model developers/owners to take responsibility for the outcomes of these scorers' scores since local programs did the training, calibration, and/or recalibration of these scorers. It is not clear what happens now at the program level to scorers whose scores are not in agreement following rescoring.

Improved selection of scorers and closer oversight of scorer training, calibration and recalibration by model developers could potentially increase the scoring reliability of candidate outcomes on the TPA. Increased oversight and/or supervision of the scorer calibration and recalibration processes could potentially involve model developers being required to provide all initial training for scorers, whether directly or through trainer of trainers individually trained and authorized by the model developer rather than through an online process only or through local teacher preparation programs with no model developer input and/or oversight.

Increased oversight and/or supervision of scorer calibration and recalibration by model developers/owners could result in higher costs for the model developers, increased need for model-specific training staff, and higher costs for programs which would be required to provide access for scorers to model-specific training. Programs that may have invested their own

resources into the scorer identification and training process may also object to increased model developer/owner control of and/or activities related to scorer training, calibration, and recalibration.

Increased model developer supervision of selection of local lead trainers

As scorer training becomes increasingly removed from the training provided directly by the model developer/owner, the likelihood of diluted quality and consistency of scorer training increases. An efficient training model often involves the use of lead trainers, who are individuals identified by local programs to train other scorers. Some models, like PACT, provide an online lead trainer component, where a model such as the CalTPA has chosen to conduct training for these individuals in person for purposes of quality control. Some TPA models hold training or other sessions periodically with the lead trainers to review their skills and training implementation processes and discuss training issues.

However, the quality of the lead trainers is critical to the continued quality of the training process and its outcomes in terms of successful training of quality local scorers. Currently, local teacher preparation programs identify the individuals whom they would like trained as lead trainers. Model developers typically do not have control over who is selected by local programs. It is a delicate situation if the model developer's staff have concerns about the quality of a given lead trainer identified by a local preparation program, all the more so because of the lack of scorer performance data that might help programs identify the best candidates for becoming a lead trainer. Since lead trainers provide initial scorer training and calibration for local teacher preparation programs, and may also provide oversight of scorer recalibration within and/or across programs, the overall quality of assessment scores within a given program hinges on the quality of the lead trainers used by that program.

One way to potentially improve the quality and consistency of lead trainers' skills could be for the model developers to have more input into the selection of the lead trainers, to require at least one in-person training session with model developer staff, and/or provide increased oversight over time for these trainers' activities. A key component is to require the lead trainers to maintain their calibration status, which because of the gravity and importance of their roles affecting the overall scoring consistency of the TPA process, should be reviewed more than once per year as is now typically done for scorers.

All of these activities would potentially have an increased cost basis for model developers/owners in terms of staff time, staff costs, and logistics for training, as well as for local preparation programs in terms of increased tracking of the performance of the lead trainers and/or in training costs for these individuals.

Specified stratified sample or random sample of rescoring

In order to obtain an appropriate sampling across candidates and scorers, the Commission could choose to specify a particular approach such as a stratified sample of candidates and scorers, or a random sample across candidates and scorers. Currently, programs are supposed to do a random sample, but it is not clear how programs are implementing that policy.

Require annual scorer data reporting in a format other than Biennial Reports

As presented in Part I.d. of this item, the current reporting related to TPA scorers does not allow for a statewide view on the reliability of scoring. The Commission could develop a scorer reliability reporting template and require all programs to submit scoring data in a consistent format on an annual basis.

Centralized rescoring (by model developers or by a contractor(s))

Because programs currently do their own rescoring, it is possible that issues of scorer calibration also affect the rescoring in the same manner as the initial scoring. In programs where there is a limited number of scorers, obtaining a true random sample of scores and of scorers may not be possible. For all programs, the current 15% rescore requirements might be modified/increased by the Commission, as outlined above.

Implementing a centralized rescoring process in collaboration with the model developers/owners' staff could result in a more accurate rescoring process across an increased number of candidates and of scorers, and thus in more reliable data about the consistency of scoring across all TPA models.

A centralized rescoring process, whether statewide or on a regional basis, could be managed:

- by TPA model developers/owners, who would select the scorers, organize the rescoring sessions, oversee the process, and analyze the rescore data for reporting to the Commission and to local programs. This approach would have cost and staffing implications for model developers/owners as well as for the Commission in its dual roles as statewide overseer of TPA implementation and CalTPA model developer/owner.
- by a contractor secured by the Commission through a competitive bid process for this function. This approach would have cost implications for the Commission in terms of preparation of a Request for Proposal (RFP), conducting a bid process, developing the contract for the successful bidder, and overseeing the contractor's work. This approach would also have cost implications for programs and/or candidates in order to fund this work through fees or other means of cost recovery.

Centralized scoring for all TPA models

Legislative guidance provides for the TPA to be embedded in local teacher preparation programs. Thus, local program scoring was established rather than a centralized scoring process that would serve the state as a whole.

There is an inherent and complex tension within the Education Code governing the TPA resulting from legislative requirements that (1) promote the development of multiple versions of an assessment that is to be locally-embedded, locally-administered, and locally-scored but that also has high stakes for candidates in that passing the assessment is one of the requirements for the recommendation for a credential, and (2) also require the assessment to provide both formative and summative outcomes information while (3) at the same time mandate each TPA assessment to demonstrate ongoing high levels of psychometric validity, scoring validity, fairness and equity for candidates as required by Commission standards, all of which are hallmarks of summative, standardized assessments that are typically centrally administered and

scored under consistent conditions rather than local assessments administered and scored under non-standard conditions.

All of the TPA models have labored to meet these somewhat contradictory expectations of local design and implementation of the assessment yet high standards of validity and reliability for administration assessment and candidate outcomes by putting into place a complex system of local coordination and oversight over the assessment process, scorer training, scorer initial calibration and continuing recalibration over time, scorer assignment and monitoring, and a double-scoring process. As a result, the TPA has become a labor-intensive assessment which adds to the overall cost, both fiscal and in terms of personnel time and effort, of locally implementing the assessment. Without such systems in place, however, the TPA would not be meeting legislative requirements for a valid and reliable candidate assessment.

The local scoring process has many established benefits, as attested to by program instructors and administrators, including providing valuable and immediate feedback for program and instructional improvement purposes. However, local scoring is a costly process for most, if not all, program sponsors. Some institutions, primarily private/independent institutions, charge students a fee that covers these costs. Some institutions pay scorers for their scoring services, while other institutions incorporate scoring into the faculty work load or make other arrangements to address their scoring needs. The cost of scoring remains a concern for all members of the TPA Users Advisory Committee.

In addition, the local scoring process exponentially increases the lack of standardization of scorer selection, scoring, training, calibration and recalibration over time. As indicated previously, this is a major factor affecting the reliability of assessment scores on the TPA.

One option could be to move to a centralized or regional scoring model for all three approved models. The Commission could issue Request for a Proposal (RFP) for a contractor to provide these services at a per-candidate cost. The per-candidate cost could be borne by the candidate, the program, or a combination. Currently-trained scorers, including faculty, field supervisors, induction support providers, master teachers and administrators, could serve as scorers through this process working with the contractor. By using trained scorers from local programs and by offering regional scoring sessions, a close link between scoring and feedback to local programs for improvement purposes could be retained. A centralized or regional scoring process operated through a contractor could provide improved scoring reliability both within and across models, programs, and individual scorers. Using centralized scoring could potentially eliminate the need for the rescoring process but might still require an auditing process. The Commission might want to consider the option of moving to a centralized or regional scoring process for all three TPA models.

Centralized administration and scoring for a single statewide model

The Education Code allows for multiple TPA models to be developed by local programs and submitted to the Commission for review and approval. However, the use of multiple TPA models makes it virtually impossible to obtain and/or analyze a statewide set of candidate data outcomes for resulting from the mandated performance assessment. As indicated above, data from multiple models administered to candidates under variable conditions and for a variable number of permitted attempts using variable scoring rubrics are not going to provide a valid or useful

statewide perspective on the effects of the performance assessment requirement. In addition, the use of multiple models that include model-specific scorer training, calibration, and recalibration increases the labor-intensity and the resulting program-level implementation costs (material and human) of each model.

Although this may not be an option that the Commission wishes to consider, the approach that would provide the maximum consistency of scoring outcomes on the TPA would be to reduce the number of available TPA models to a single statewide model, whether this model were to be locally implemented and scored, or centrally or perhaps regionally scored. This process would both increase to the maximum level possible the scoring reliability for the TPA and eliminate the need for the rescoring process.

For this to occur, the developers of the current three models might be encouraged or facilitated to work together to develop a single model that incorporates the best features of each model into a single assessment design. Alternatively, the nationally available TPAC could be evaluated for this purpose. As the national climate of teacher preparation has shifted recently toward a growing interest in performance-based measures of teacher candidates, states are looking for available options for performance assessments and TPAC has been working with states to address that need (Appendix C).

A Final Consideration

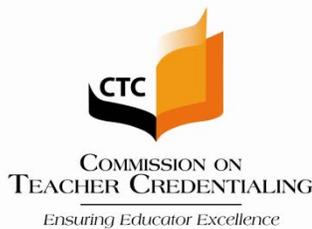
Like all of the Commission's examinations, the TPA is a large-scale assessment that has stakes for candidates. Two sets of standards affect how the TPA is designed and implemented, regardless of model. The first set is the Commission's Assessment Design Standards (Appendix B). The two Assessment Design Standards were written at a time when it was expected that the TPA would be a standardized assessment that would be centrally administered and scored. The Commission has no data at the present time regarding how the model developers/owners have continued to meet these standards over time. In addition, model developers may have insufficient resources for ongoing updating of the assessment and for continuous data collection and analysis as required by the current Design Standards, whereas a contractor conducting a centralized administration and scoring process would have a larger staff and fiscal resources resulting typically from candidate fees for the assessment. Reviewing and revising the Commission's Assessment Design Standards is another TPA-related priority for future Commission consideration and potential action.

As a large scale state-mandated assessment, however, the TPA is also subject to the assessment quality standards represented by the *Joint Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council for Measurement in Education. These standards clearly outline requirements for assessment reliability, as well as for many other psychometric properties and requirements, that all three TPA models should meet. These standards are designed to assure that the properties of assessments that contribute to decisions about individual candidates are legally defensible. The considerations discussed above for potential Commission action regarding the TPA have been formulated with the *Joint Standards* in mind.

Next Steps and Future Agenda Items

Based on Commission discussion and direction, staff will develop and present future agenda items related to the teaching performance assessment for Commission review and potential action.

Appendix A



PROGRAM SPONSOR ALERT

Date: August 12, 2010

Number: 10-17

Subject: Accreditation Processes Related to the Implementation of the Teaching Performance Assessment (TPA)

Summary

The Committee on Accreditation (COA) and the Teaching Performance Assessment Users Advisory Committee (UAC), a statewide oversight group representing the three Commission-approved TPA models, met several times recently to discuss how the accreditation system provides oversight to TPA implementation for Multiple and Single Subject teacher preparation programs. On August 4, 2010 the Committee on Accreditation approved several refinements to the accreditation system with respect to the TPA and MS/SS preparation programs. The refinements impact all major activities of the accreditation system.

1. Biennial Reports: Scorer data will be submitted
2. Program Assessment: Review process for Standards 17-19
3. Site Visits: Resources are being developed for use at the site visit

This Program Sponsor Alert describes the refinements.

Background

The Teaching Performance Assessment (TPA) has been a requirement for all Preliminary Multiple and Single Subject candidates admitted to a teacher preparation program on or after July 1, 2008. There currently are three Commission-approved models: the CalTPA, Performance Assessment for California Teachers (PACT), and Fresno Assessment of Student Teachers (FAST). All three models have some commonalities such as specific tasks that candidates must accomplish, an extensive scorer training system, and rubric scoring based on a four-point scale. In addition, each model has requirements and processes that distinguish it from the other two models.

Three standards apply to how a program implements its chosen TPA-model that are reviewed during the accreditation activities. Specifically, the accreditation process is charged with providing oversight

of the TPA implementation process. The standards that apply to the implementation of the TPA are contained in Category E: Standards 17-19 below.

Standard 17: Implementation of the Teaching Performance Assessment (TPA): Program Administration Processes

Standard 18: Implementation of the Teaching Performance Assessment Candidate Preparation and Support

Standard 19: Implementation of the Teaching Performance Scorer Qualifications, Training and Scoring Reliability

Changes to the Biennial Report Data Requirement for Multiple and Single Subject Programs

The UAC and the COA discussed at length the role that scorer information plays in understanding whether a program is meeting the implementation standards for the teaching performance assessment. Program Standard 19 states:

The program provides assessor training and/or facilitates assessor access to training in the specific TPA model(s) used by the program. The program selects assessors who meet the established selection criteria and uses only assessors who successfully complete the required TPA model assessor training sequence and who have demonstrated initial calibration to score candidate TPA responses.

The program periodically reviews the performance of assessors to assure consistency, accuracy, and fairness to candidates within the TPA process, and provides recalibration opportunities for assessors whose performance indicates they are not providing accurate, consistent, and/or fair scores for candidate responses.

The program complies with the assessor recalibration policies and activities specific to each approved TPA model, including but not limited to at least annual recalibration for all assessors, and uses and retains only TPA assessors who consistently maintain their status as qualified, calibrated, program-sponsored assessors. The program monitors score reliability through a double-scoring process applied to at least 15% of TPA candidate responses.

The COA and UAC agreed that information related to scorer training and calibration is critical contextual information for understanding how the teaching performance assessment is being implemented in each MS and SS program.

To that end, the COA approved revisions to the biennial report requirements that will capture information about scorers, such as training and (re)calibration, in the implementation of the TPA. The additional information now required to be submitted in the biennial reports for Multiple and Single Subject programs is the following:

- 1) Number of Scorers: The total number of scorers the program uses and the number of scorers who scored in the years for which the biennial report data is being submitted.
- 2) Scorer Initial Training and Recalibration: The number of scorers who successfully completed initial training and the number who recalibrated for the applicable biennial report years.
- 3) Data on Reliability Related to Double Scoring (% of score agreement).

- 4) Modifications made to scorer selection, training, recalibration. This information may be included in Section A, Part I or in Section A, Part IV.

For those submitting in Fall 2010, this additional information is voluntary, but highly encouraged. This information may be included in aggregated data (preferable) or in narrative form. Those institutions submitting reports in August 2010 may submit an addendum with this information any time prior to December 15, 2010. The UAC and COA will review the types of information submitted this Fall and may provide additional guidance to the multiple and single subject programs as to best practices in submitting scorer data in future Biennial Reports.

Biennial reports due in Fall 2011 must include the data identified in 1-3 above, as well as information on 4 above, for Multiple and Single Subject teacher preparation programs.

The Biennial Report Template has been revised and is available on the website: <http://www.ctc.ca.gov/educator-prep/program-accred-biennial-reports.html>.

Changes in the Program Assessment Review of Standards 17-19

Each sponsor's implementation of program standards is reviewed via an in-depth document review during Program Assessment. Training all BIR members to understand the highly technical implementation requirements for each of the TPA models and of Standards 17-19 poses a significant challenge for the Commission. However, review of the program responses to these standards requires that reviewers have a deep understanding of the three approved TPA models. Therefore, the UAC suggested and the COA agreed on a modification to the review process during Program Assessment of these three TPA-focused standards.

Rather than expecting every program assessment reviewer to review all standard responses, including Standards 17-19, submitted by Multiple or Single Subject programs, a subset of BIR reviewers with particular expertise in the TPA will review the responses to Standards 17-19. Other BIR team members will focus their review of the responses to Standards 1-16. This will ensure a fair and rigorous process for the review of Standards 17-19 regardless of TPA model. It will also allow those with expertise in the variations of delivery of particular models to accurately assess whether the TPA is being implemented in accordance with the model as required by Standard 17. The *Preliminary Findings of Program Assessment* reviewers will still be confirmed through interviews and the review of other evidence by BIR members at the site visit.

To ensure that Program Assessment readers provide consistent reviews across models, institutions, and credential pathways, the TPA Users Group and the COA developed a list of guiding questions (Appendix A). These questions are not intended to replace the TPA related standards, but rather to guide Program Assessment readers to ask critical, but uniform questions of each program's response that help determine whether a program is meeting Commission adopted standards. Institutions preparing responses may also find these questions helpful as they prepare program assessment documents, but the institution's response needs to meet the language of the adopted standards.

Changes to the Site Visit Review of Standards 17-19

No substantive changes to the manner in which the site visit team reviews Standards 17-19 will take place at this time. However, the UAC suggested and the COA approved the development of additional resources to assist site visit teams in their review of Standards 17-19, including the last column of the table in Appendix A that identifies the individuals most likely to have the information

necessary for reviewing the implementation of Standards 17-19 (See Appendix A). In addition, a brief synopsis of each of the three approved models for the TPA will be provided to site visit team members.

The UAC and the COA will continue to monitor the process through which TPA implementation is reviewed in the Commission's accreditation activities.

References

COA Agenda Items

- June 2010 <http://www.ctc.ca.gov/educator-prep/coa-agendas/2010-06/2010-06-item-16.pdf>
- Insert for June 2010 <http://www.ctc.ca.gov/educator-prep/coa-agendas/2010-06/2010-06-item-16-insert.pdf>
- August 2010 <http://www.ctc.ca.gov/educator-prep/coa-agendas/2010-08/2010-08-item-17.pdf>

Contact Information

For additional information on this topic, contact BiennialReports@ctc.ca.gov.

Standards 17-19
Considerations for Program Assessment and Site Visit

Adopted Standard	Program Assessment Considerations	Site Visit Considerations*
Standard 17: Implementation of the Teaching Performance Assessment (TPA): Program Administration Processes		
<p>The TPA is implemented according to the requirements of the Commission-approved model selected by the program.* One or more individuals responsible for implementing the TPA document the administration, scoring, and data reporting processes for all tasks/activities of the applicable TPA model in accordance with the requirements of the selected model.</p>	<ol style="list-style-type: none"> 1. Does the response clearly indicate that the TPA is implemented according to the Commission-approved model selected by the program? – <i>Hold answering this question until all other aspects of the TPA related standards have been reviewed.</i> 2. Does the response clearly indicate who is responsible for the implementation of the TPA including? <ol style="list-style-type: none"> a. Administration b. Scoring c. Data reporting 	<p>Administrators (program) Assessment Coordinators Credential Analyst Data Analyst Faculty Lead Scorers Program Coordinator Staff TPA Coordinator</p>
<p>The program adopts a passing score standard and provides a rationale for establishing that passing standard.</p>	<ol style="list-style-type: none"> 3. Does the response clearly state the passing score standard adopted and the rationale for the passing score? 	<p>Administrators (program) Assessment Coordinators Faculty Program Coordinator TPA Coordinator</p>
<p>The program maintains both program level and candidate level TPA data, including but not limited to individual and aggregated results of candidate performance, assessor calibration status, and assessor performance over time.</p>	<ol style="list-style-type: none"> 4. Does the response clearly indicate how the program collects and maintains program level and candidate level data? <ol style="list-style-type: none"> a) Individual candidate performance results b) Aggregated candidate performance results c) Assessor calibration status d) Assessor performance over time 	<p>Administrators (program) Assessment Coordinators Credential Analyst Data Analyst Program Coordinator Staff TPA Coordinator</p>

Adopted Standard	Program Assessment Considerations	Site Visit Considerations*
<p>The program documents the use of these data not only for Commission reporting and/or accreditation purposes, but also for program improvement.</p>	<p>5. Does the response clearly indicate how the data are being used to reflect on the program and used for program improvement?</p>	<p>Administrators (program) Assessment Coordinators Data Analyst Faculty Program Coordinator TPA Coordinator University Supervisors</p>
<p>The program assures that candidates understand the appropriate use of their performance data as well as privacy considerations relating to candidate data.</p> <p>The program also consistently uses appropriate measures and maintains documentation to assure the privacy of the candidate, the K-12 students, the school site and school district, and other adults involved in the TPA process.</p> <p>The program establishes and consistently uses appropriate measures to ensure the security of all TPA materials, including all print, online, video, candidate, and assessor materials.</p>	<p>6. Does the response clearly indicate processes and policies relevant to the following:</p> <ul style="list-style-type: none"> a) Informing candidates about appropriate use of data b) Protecting candidate privacy c) Protecting the privacy of K-12 students, school site, and school district, and other adults involved in the TPA process. d) how candidates are informed of the appropriate uses of their performance data and the privacy of candidates and candidate data? e) Does the process clearly describe the process to ensure the security of all TPA materials? 	<p>Administrators (program) Assessment Coordinator Candidates Credential Analyst Data Analyst District Based Supervisors Faculty Graduates Lead Assessors Program Coordinator TPA Coordinator University Based Field Supervisors</p>

Adopted Standard	Program Assessment Considerations	Site Visit Considerations*
Standard 18: Implementation of the Teaching Performance Assessment: Candidate Preparation and Support		
<p>The teacher preparation program assures that each candidate receives clear and accurate information about the nature of the pedagogical tasks within the Commission-approved teaching performance assessment model selected by the program, the passing score standard adopted by the program, and the opportunities available within the program to prepare for completing the TPA tasks/activities.</p> <p>The program assures that candidates understand that all responses to the TPA that are submitted for scoring must represent the candidate's own unaided work.</p> <p>The program assures that candidates understand and follow the appropriate policies and procedures to protect the privacy and confidentiality of the K-12 students, teachers, school sites, school districts, adults, and others who are involved in any of the components of the TPA tasks/activities.</p>	<p>1. Does the response clearly indicate how the program communicates its particular implementation strategy and requirements to the candidates including?</p> <ul style="list-style-type: none"> a) passing score standard b) opportunities within the program to prepare for completing the TPA tasks/activities c) that work scored is unaided candidate work d) appropriate policies and procedures to protect privacy and confidentiality of the K-12 students, teachers, school sites, school districts, adults, and others who are involved in any components of the TPA. 	<p>Administrators (program, and employers) Assessment Coordinators Candidates District Based Supervisors Faculty Graduates Lead Assessors Program Coordinator TPA Coordinator University Based Field Supervisors</p>

Adopted Standard	Program Assessment Considerations	Site Visit Considerations*
Standard 19: Implementation of the Teaching Performance: Assessor Qualifications, Training, and Scoring Reliability		
<p>The teacher preparation program establishes selection criteria for assessors of candidate responses to the TPA. The selection criteria include but are not limited to pedagogical expertise in the content areas assessed within the TPA.</p> <p>The program provides assessor training and/or facilitates assessor access to training in the specific TPA model(s) used by the program.</p> <p>The program selects assessors who meet the established selection criteria and uses only assessors who successfully complete the required TPA model assessor training sequence and who have demonstrated initial calibration to score candidate TPA responses.</p>	<ol style="list-style-type: none"> 1. Does the response clearly indicate the selection criteria for TPA assessors and that they document that assessors meet the selection criteria? 2. Does the response clearly indicate how the program provides the assessor training process? 3. Does the response clearly indicate how the program documents successful completion of assessor training for all assessors? 	<p>Administrators (program) Assessment Coordinators Assessors Lead Assessors Program Coordinator TPA Coordinator</p>
<p>The program periodically reviews the performance of assessors to assure consistency, accuracy, and fairness to candidates within the TPA process, and provides recalibration opportunities for assessors whose performance indicates they are not providing accurate, consistent, and/or fair scores for candidate responses.</p> <p>The program complies with the assessor recalibration policies and activities specific to each approved TPA model, including but not limited to at least annual recalibration for all assessors, and uses and retains only TPA assessors who consistently maintain their status as qualified, calibrated, program-sponsored assessors.</p>	<ol style="list-style-type: none"> 4. Does the response clearly describe the programs recalibration policies and processes including: <ol style="list-style-type: none"> a) how the program periodically reviews assessor performance, b) identify assessors who are in need of recalibration, and the program provides those additional training opportunities? and c) Annual recalibration for all assessors 	<p>Administrators (program) Assessment Coordinators Assessors Lead Assessors Program Coordinator TPA Coordinator</p>

Adopted Standard	Program Assessment Considerations	Site Visit Considerations*
Standard 19: Implementation of the Teaching Performance: Assessor Qualifications, Training, and Scoring Reliability		
<p>The program monitors score reliability through a double-scoring process applied to at least 15% of TPA candidate responses.</p>	<p>5. Does the response clearly indicate how the program monitors score reliability and a double-scoring process applied to at least 15% of candidate responses?</p>	<p>Administrators (program) Assessment Coordinators Assessor Lead Assessors Program Coordinator TPA Coordinator</p>
<p>The program establishes and maintains policies and procedures to assure the privacy of assessors as well as of information about assessor scoring reliability.</p>	<p>6. Does the response clearly describe the policies and procedures to assure the privacy of assessors?</p>	<p>Administrators (program) Assessment Coordinators Assessors Lead Assessors Program Coordinator TPA Coordinator</p>
<p>In addition, the program maintains the security of assessor training materials and protocols in the event that the program uses its own assessors (such as, for example, a designated Lead Assessor) to provide local assessor training.</p>	<p>7. If applicable, does the response clearly describe how the program maintains the privacy of assessor materials?</p>	<p>Administrators (program) Assessment Coordinators Assessors Lead Assessors Program Coordinator TPA Coordinator</p>

Appendix B

Assessment Design Standard 1: Assessment Designed for Validity and Fairness (Assessment Design Standard 1 Applies to Programs that Request Approval of Alternative Assessments)

The sponsor of the professional teacher preparation program requests approval of a Teaching Performance Assessment (TPA) in which complex pedagogical assessment tasks and multi-level scoring scales are linked to the Teaching Performance Expectations (TPEs). The program sponsor clearly states the intended uses of the assessment, anticipates its potential misuses, and ensures that local uses are consistent with the statement of intent. The sponsor maximizes the fairness of assessment design for all groups of candidates in the program, and ensures that the established passing standard on the TPA is equivalent to or more rigorous than the recommended state passing standard.

Required Elements for Assessment Design Standard 1: Assessment Designed for Validity and Fairness

- 1(a) The Teaching Performance Assessment includes complex pedagogical assessment tasks to prompt aspects of candidate performance that measure the TPEs. Each task is substantively related to two or more major domains of the TPEs. For use in judging candidate-generated responses to each pedagogical task, the assessment also includes multi-level scoring scales that are clearly related to the same TPEs that the task measures. Each task and its associated scales measure two or more TPEs. Collectively, the tasks and scales in the assessment address key aspects of the six major domains of the TPEs. The sponsor of the professional teacher preparation program documents the relationships between TPEs, tasks and scales.
- 1(b) To preserve the validity and fairness of the assessment over time, the sponsor may need to develop and field-test new pedagogical assessment tasks and multi-level scoring scales to replace or strengthen prior ones. Initially and periodically, the sponsor analyzes the assessment tasks and scoring scales to ensure that they yield important evidence that represents candidate knowledge and skill related to the TPEs, and serves as a basis for determining entry-level pedagogical competence to teach the curriculum and student population of California's K-12 public schools. The sponsor records the basis and results of each analysis, and modifies the tasks and scales as needed.
- 1(c) Consistent with the language of the TPEs, the sponsor defines scoring scales so different candidates for credentials can earn acceptable scores on the Teaching Performance Assessment with the use of different pedagogical practices that support implementation of the K-12 content standards and curriculum frameworks. The sponsor takes steps to plan and anticipate the appropriate scoring of candidates who use pedagogical practices that are educationally effective but not explicitly anticipated in the scoring scales.
- 1(d) The sponsor develops scoring scales and assessor training procedures that focus primarily on teaching performance and that minimize the effects of candidate factors that are not clearly related to pedagogical competence, which may include (depending on the circumstances) factors such as personal attire, appearance, demeanor, speech patterns and accents that are not likely to affect student learning.

- 1(e) The sponsor publishes a clear statement of the intended uses of the assessment. The statement demonstrates the sponsor's clear understanding of the high-stakes implications of the assessment for candidates, the public schools, and K-12 students. The statement includes appropriate cautions about additional or alternative uses for which the assessment is not valid. Before releasing information about the assessment design to another organization, the sponsor informs the organization that the assessment is valid only for determining the pedagogical competence of candidates for initial teaching credentials in California. All elements of assessment design and development are consistent with the intended use of the assessment for determining the pedagogical competence of candidates for Preliminary Teaching Credentials in California.
- 1(f) The sponsor completes content review and editing procedures to ensure that pedagogical assessment tasks and directions to candidates are culturally and linguistically sensitive, fair and appropriate for candidates from diverse backgrounds. The sponsor ensures that groups of candidates interpret the pedagogical tasks and the assessment directions as intended by the designers, and that assessment results are consistently reliable for each major group of candidates.
- 1(g) The sponsor completes basic psychometric analyses to identify pedagogical assessment tasks and/or scoring scales that show differential effects in relation to candidates' race, ethnicity, language, gender or disability. When group pass-rate differences are found, the sponsor investigates to determine whether the differences are attributable to (a) inadequate representation of the TPEs in the pedagogical tasks and/or scoring scales, or (b) overrepresentation of irrelevant skills, knowledge or abilities in the tasks/scales. The sponsor acts promptly to maximize the fairness of the assessment for all groups of candidates and documents the analysis process, findings, and action taken.
- 1(h) In designing assessment administration procedures, the sponsor includes administrative accommodations that preserve assessment validity while addressing issues of access for candidates with disabilities.
- 1(i) In the course of developing or adopting a passing standard that is demonstrably equivalent to or more rigorous than the State recommended standard, the sponsor secures and reflects on the considered judgments of teachers, the supervisors of teachers, the support providers of new teachers, and other preparers of teachers regarding necessary and acceptable levels of proficiency on the part of entry-level teachers. The sponsor periodically reconsiders the reasonableness of the scoring scales and established passing standard.

Assessment Design Standard 2: Assessment Designed for Reliability and Fairness
(Assessment Design Standard 2 Applies to Programs that Request Approval of Alternative Assessments)

The sponsor of the professional teacher preparation program requests approval of an assessment that will yield, in relation to the key aspects of the major domains of the TPEs, enough collective evidence of each candidate's pedagogical performance to serve as an adequate basis to judge the candidate's general pedagogical competence for a Preliminary Teaching Credential. The sponsor carefully monitors assessment development to ensure consistency with the stated purpose of the assessment. The Teaching Performance Assessment includes a comprehensive program to train and re-train assessors. The sponsor periodically evaluates assessment design to ensure equitable treatment of candidates. The assessment design and its implementation contribute to local and statewide consistency in the assessment of teaching competence.

Required Elements for Assessment Design Standard 2: Assessment Designed for Reliability and Fairness

- 2(a) In relation to the key aspects of the major domains of the TPEs, the pedagogical assessment tasks and the associated directions to candidates are designed to yield enough evidence for an overall judgment of each candidate's pedagogical qualifications for a Preliminary Teaching Credential. The program sponsor will document sufficiency of candidate performance evidence through thorough field-testing of pedagogical tasks, scoring scales, and directions to candidates.
- 2(b) Pedagogical assessment tasks and scoring scales are extensively field-tested in practice before being used operationally in the Teaching Performance Assessment. The sponsor of the program evaluates the field-test results thoroughly and documents the field-test design, participation, methods, results and interpretation.
- 2(c) The Teaching Performance Assessment system includes a comprehensive program to train assessors who will score candidate responses to the pedagogical assessment tasks. An assessor training pilot program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and the multi-level scoring scales. The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy in relation to the scoring scales associated with the task. When new pedagogical tasks and scoring scales are incorporated into the assessment, the sponsor provides additional training to the assessors, as needed.
- 2(d) In conjunction with the provisions of Teacher Preparation Program Standard 19, the sponsor plans and implements periodic evaluations of the assessor training program, which include systematic feedback from assessors and assessment trainers, and which lead to substantive improvements in the training as needed.
- 2(e) The program sponsor requests approval of a detailed plan for the scoring of selected assessment tasks by two trained assessors for the purpose of evaluating the reliability of scorers during field-testing and operational administration of the assessment. The subsequent assignment of one or two assessors to each assessment task is based on a cautious interpretation of the ongoing evaluation findings.

- 2(f) The sponsor carefully plans successive administrations of the assessment to ensure consistency in elements that contribute to the reliability of scores and the accurate determination of each candidate's passing status, including consistency in the difficulty of pedagogical assessment tasks, levels of teaching proficiency that are reflected in the multilevel scoring scales, and the overall level of performance required by the Commission's recommended passing standard on the assessment.
- 2(g) The sponsor ensures equivalent scoring across successive administrations of the assessment and between the Commission's model and local assessments by: using marker performances to facilitate the training of first-time assessors and the further training of continuing assessors; monitoring and recalibrating local scoring through third party reviews of scores that have been assigned to candidate responses; and periodically studying proficiency levels reflected in the adopted passing standard.
- 2(h) The sponsor investigates and documents the consistency of scores among and across assessors and across successive administrations of the assessment, with particular focus on the reliability of scores at and near the adopted passing standard. To ensure that the overall construct being assessed is cohesive, the sponsor demonstrates that scores on each pedagogical task are sufficiently correlated with overall scores on the remaining tasks in the assessment. The sponsor demonstrates that the assessment procedures, taken as a whole, maximize the accurate determination of each candidate's overall pass-fail status on the assessment.
- 2(i) The sponsor's assessment design includes an appeal procedure for candidates who do not pass the assessment, including an equitable process for rescoring of evidence already submitted by an appellant candidate in the program.

Appendix C

Update on TPAC

The following information regarding the development and current status of the national TPAC effort comes from the public AACTE (American Association of Colleges of Teacher Education) website: <http://aacte.org/Programs/Teacher-Performance-Assessment-Consortium-TPAC/teacher-performance-assessment-consortium.html> (November 2011)

One of the few areas of consensus among education policy makers, practitioners and the general public today is that improving teacher quality is one of the most direct and promising strategies for improving public education outcomes in the United States. However, existing federal, state, and local policies for defining and measuring teacher quality rely almost exclusively on classroom observations by principals that differentiate little among teachers and offer little useful feedback, on the one hand, or teachers' course-taking records plus paper-and-pencil tests of basic academic skills and disciplinary subject matter knowledge that are poor predictors of later effectiveness in the classroom, on the other. It has become clear that new strategies for evaluating teacher competence and effectiveness are needed.

The American Association of Colleges of Teacher Education (AACTE) and Stanford University have formed a partnership to develop the Teacher Performance Assessment (TPA), a 21-state initiative involving over 100 teacher preparation programs. The Teacher Performance Assessment will create a body of evidence of teaching competence, providing a vehicle for systematically examining the assessment data to improve teacher preparation programs, provide professional development to practicing teachers and inform decisions about tenure of individual teachers.

This instrument, based on the highly successful Performance Assessment for California Teachers (PACT), will be made available to states and teacher preparation programs that wish to improve the consistency with which teacher licensure and accreditation decisions are made, including the rapidly expanding number and variety of "alternative routes" to licensure. It will also be available for use by states and their school districts to evaluate and inform continuation-of -employment decisions about teachers already practicing in their classrooms.

The assessment system consists of two components: 1) Embedded Signature Assessments (ESAs) that vary across programs; and 2) a common portfolio assessment, and the Teaching Event. The ESAs are formative signature assignments embedded in coursework. The ESAs vary across programs, are mission driven and reflect program-specific teaching philosophies or goals that contribute to the unique character of program graduates. For example, embedded assessments may include child case studies, planning instructional units, analyses of student work, and observations of student teaching.

The Teacher Performance Assessment consists primarily of a series of Teaching Events, a multiple measure assessment system documenting teaching and learning in 3-5 day learning segments for one class of students. Teaching Events are subject-specific, with

separate forms for Multiple Subject (elementary) and Single Subject (secondary) credential areas. The specific records of practice (evidence) in the Teaching Event consist of artifacts of teaching (lesson plans, video clips of instruction, student work samples, teacher assignments, daily reflections) and reflective commentaries which explain the professional judgments underlying the teaching and learning artifacts.

Development of a nationally accessible teaching performance assessment will allow states, school districts and teacher preparation programs to share a common framework for defining, and measuring a set of core teaching skills that form a valid and robust vision of teacher competence. As states reference data generated from this tool to inform teacher licensure, recruitment and tenure, they will establish a national standard for relevant and rigorous practice that advances student learning.

TPA Goals:

- Improve student outcomes
- Improve the information base guiding improvement of teacher preparation programs
- Strengthen the information base for accreditation and comparison of program effectiveness
- Be used in combination with other measures as a requirement for licensure
- Guide professional development for teachers across the career continuum
- Serve as a model for assessments, sitting in between the assessment for initial licensure and National Board certification, e.g., continuation-of-employment, tenure, and career ladder decisions.

Current Project Status

- Eleven states participated in the spring 2010 tryouts designed to give institutes of higher education (IHEs) some experience with the instrument before we began refining the instrument for the pilot.
- TPAC's Design Team met in July to address feedback supplied by candidate and faculty members who tried out tasks in the TPA instrument during spring 2010. In direct response to these reviews, changes were implemented for the final draft assessment by the Stanford team.
- The first meeting of the newly established TPAC Advisory Council took place on June 28. The Council reviewed key aspects of the project, including the policy agenda, the communications plan, TPA research, and funding status, with the goal of obtaining solid advice and support in the development of the project.
- Massachusetts, Minnesota, Ohio, Tennessee, and Washington are accelerating their participation in the project by including all of their IHEs in the field test next year, due to the expectation that their states will allow or require the use of TPA in licensure, accreditation, and/or certification as early as 2012.
- In Spring 2011, programs began piloting assessments in eight areas: elementary literacy and mathematics, secondary English-language arts, history-social science, mathematics, and science; special education and early childhood special education, and early childhood.
- Secured commitments from 24 participating pilot states, consisting of teams made up of representatives from state education agencies (SEAs) and over 100 teacher preparation

institutions, and conducted a face-to-face meeting to ready the states for implementation of the 3-year pilot. The 24 states include:

California	Iowa	Missouri	Oregon	Tennessee
Colorado	Maryland	New Jersey	Virginia	
Delaware	Massachusetts	New York	Washington	
Georgia	Michigan	North Carolina	West Virginia	
Idaho	Minnesota	Ohio	Wisconsin	
Illinois	Oklahoma	Wyoming	District of Columbia	

In addition, Western Governors University (WGU) is participating in the pilot. WGU is an online accredited teacher preparation program in 49 states.

Appendix D

Executive Summary of the Options Presented in this Item

Increasing the Reliability (Scoring Consistency) of Candidate Outcomes on the Teaching Performance Assessment (TPA)

Option For Increasing Reliability of Candidate Scores	Implications		
	Approved Programs	Model Developers (CalTPA, PACT, FAST)	Commission
Establishment of a minimally acceptable rate of scorer agreement to verify scoring consistency (within current 15% rescore).	Potential changes in scorers/scoring practices if programs find scorers are not meeting the minimally acceptable rate.	Potential establishment of minimally acceptable rate by model, additional support for programs.	Develop policy regarding what to do if programs are not meeting a minimally acceptable rate of scorer agreement; potentially modify reporting requirement(s).
Increased minimum number and/or percentage of rescors, within each program.	Increased number of minimum rescors for all programs; larger impact on smaller programs; additional scorer workload/costs.	Potential additional training for lead scorers to meet increased scoring demand	Policy updates and/or statutory change, communication to the field. Increased training costs for CalTPA.
Increased frequency of required scorer recalibration.	Increased scoring costs for more frequent scorer recalibration and scoring management.	Increased updates of online recalibration systems.	Policy/Standards updates, communication to the field.
Increased model developer supervision of scorer calibration/recalibration.	Less autonomy for recalibrating scorers, likely increased costs for recalibration of scorers.	Increased oversight of what programs are doing with scorers, increased costs for additional recalibration options.	Policy/Standards updates, communication to the field; increased workload for CalTPA staff.
Increased model developer supervision of selection of local lead trainers.	Less autonomy for training scorers, likely increased costs for training of scorers.	Increased oversight of what programs are doing with scorers, increased costs for additional scorer training. No impact for FAST.	Policy/Standard updates, communication to the field; increased workload for CalTPA staff
Specified stratified random sample of rescors.	Potential changes in practice of selecting candidates for re-score.	Updated implementation guides, increased support for programs.	Policy/Standards updates, communication to the field.
Require annual scorer data reporting in a format other than Biennial Reports.	Additional data template to complete and submit annually (similar to candidate data template).	Increased support for programs.	Develop templates, process for distribution and collection, integration with other data systems, increased staff workload to maintain, compile and analyze data and produce reports.

Increasing the Reliability (Scoring Consistency) of Candidate Outcomes on the Teaching Performance Assessment (TPA)

Option For Increasing Reliability of Candidate Scores	Implications		
	Approved Programs	Model Developers (CalTPA, PACT, FAST)	Commission
Centralized rescoring by model on a statewide or regional basis: -by model developers -by a contractor	Potential reduced scoring costs (unless programs or candidates were charged a fee). Programs would still contribute scorers for the process.	Increased costs/work/staff for taking on re-scoring, unless work organized by a contractor and overseen by staff.	Potential development of an RFP, selection of a contractor, and increased staff costs for overseeing contractor's work. Assumes no-cost contract based on charging a fee for assessment scoring. Could potentially be managed by model developers but dependent on funding and staff availability.
Centralized scoring for all models.	Potential reduced scoring costs, reduced workload for program staff. Also potential decrease in professional development (where faculty score TPA). Potential fees for TPA scoring to be borne by programs or candidates. Programs would still contribute scorers for the process.	Potential increased participation in the scoring process, working with a contractor to ensure each model is scored correctly and effectively.	Development of an RFP, selection of a contractor, increased staff costs for overseeing contractor's work, working with programs to determine best sources of funding assuming work is done under a no-cost contract, work with models to develop scoring procedures, policy changes for contractor implementation.
Centralized administration and scoring for a single statewide model.	Reduced local control of TPA model selection and administration. Elimination of scoring tasks for programs. Potential fees for scoring to be borne by programs and/or candidates. Programs would still contribute scorers for the process.	Only a single model would be implemented; other models would be discontinued.	Policy changes and/or statutory change. Need to develop/select a single statewide TPA model. Development of an RFP, selection of a contractor for single statewide TPA model. Development of an RFP, selection of a contractor for organizing the scoring of the single statewide TPA model. Oversight of contractor's work during implementation.