2019 No. 120

An Investigation of the Comparability of Commission-Approved Teaching Performance Assessment Models

Final Report – Volume I: Technical Report

Prepared for: California Commission on Teacher Credentialing
1900 Capitol Avenue
Sacramento, CA 95811
Attn: Michael Taylor

Editors: Andrea L. Sinclair
Arthur Thacker

Date: December 31, 2019

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | humrro.org

Acknowledgments

First and foremost, we wish to thank the Technical Advisory Committee for their time, insight, and guidance throughout the project: Jean Behrend, Thomas Guskey, Helene Mandell, Nicole Merino, Raymond Pecheone, Amy Reising, and Lauress (Laurie) Wise. In addition, we are extremely grateful for the time and energy that the model representatives devoted to ensuring that we had the documentation and materials necessary to conduct the investigation. Heather Klesch served double duty in this regard, suppling vital information for the CalTPA and edTPA models. We also extend our gratitude to the leadership team at the California Commission on Teacher Credentialing, notably Mary Vixie Sandy and Teri Clark, for supporting this important and meaningful research. Last, but certainly not least, we thank Michael Taylor and Kathryn Taylor who served as our primary contacts on this investigation for the Commission, and who provided timely and helpful guidance from start to finish. The guidance and support each of you provided to this work reflects deep expertise, as well as an inspiring commitment to the teaching profession.

An Investigation of the Comparability of Commission-Approved Teaching Performance Assessment Models: Final Report

Executive Summary

California's Commission on Teacher Credentialing (Commission) requires all programs of preliminary multiple and single subject teacher preparation to use a Commission-approved Teaching Performance Assessment (TPA) as one of the program completion requirements for prospective teacher candidates. Three TPA models were approved by the Commission. Each is listed below along with its model sponsor.[1]

- the FAST (Fresno Assessment of Student Teachers), owned and operated by California State University, Fresno (Fresno State);
- the edTPA, owned by Stanford University, with an operational contractor of the Evaluation Systems group of Pearson; and
- the CalTPA (California Teaching Performance Assessment), originally developed by Educational Testing Service (ETS) and owned by the Commission, revised by a Design Team with an operational contractor, also of the Evaluation Systems group of Pearson.

The Commission adopted revised *Assessment Design Standards* (ADS) in December 2015 and *Teaching Performance Expectations* (TPEs) in June 2016. The ADS describe the design requirements for all TPA models and the TPEs describe the performance standards for beginning teachers. There are six TPE "domains" and each domain includes six to eight descriptors, referred to as "elements," which describe the knowledge, skills, and abilities (KSAs) required for beginning teachers. Each TPA model must adhere to the ADS and measure the TPEs.

Although each TPA model is Commission-approved, they differ in several important ways—for example, in the design of candidate tasks or cycles and scoring rubrics. Further, each TPA model must address all TPE domains but need not address all elements. These inter-model differences raise questions regarding the comparability of results obtained by teacher candidates completing the various TPAs. Consequently, the Commission contracted with the Human Resources Research Organization (HumRRO) to conduct an independent investigation of the comparability of the three TPA models. The investigation occurred between June 2017 and December 2019.

The technical approach to the evaluation marshalled evidence from numerous sources such as stakeholder surveys, analysis of score patterns, and comparisons to a common criterion. The goal was to accumulate as much evidence as possible (i.e., a "body of evidence") to evaluate the comparability of the three TPA models.

This investigation adopted a "Theory of Action approach" (Kane, 2006; 2013) to identify the claims that need to be substantiated to "assure that the Commission-approved TPA models are *sufficiently comparable* [emphasis added] that they are equitably assessing candidates working toward a California preliminary multiple or single subject teaching credential" (Request for Proposal, p. 5). This investigation was guided by a technical advisory committee (TAC)

[1] Per the *Assessment Design Standards,* "model sponsor" refers to the entity that represents the assessment and is responsible to programs using that model and to the Commission.

composed of model sponsors and independent assessment experts. During the first TAC meeting the attendees discussed the meaning of "sufficiently comparable." This discussion resulted in the following guidance: "comparable does not mean that the models are equal in *how* they measure the KSAs required by the TPEs, but that all models equitably identify TPE-ready professionals." To assure that this ultimate objective is attained the Theory of Action requires the following claims be substantiated:[2]

- Claim 1: The TPA models are sufficiently comparable in their representation of the Commission's *Assessment Design Standards* (ADS) and in their assessment and weighting of the Commission-adopted *Teaching Performance Expectations* (TPEs).
- Claim 2: The guidance and supports (e.g., guide/manual/handbook and other resources) provided by model sponsors to candidates and teacher preparation faculty are sufficiently clear and detailed to ensure that the model is implemented as designed and intended.
- Claim 3: The scoring rubrics for each TPA model are sufficiently clear and detailed to ensure that trained scorers can accurately and consistently score candidate submissions.
- Claim 4: For each TPA model, there is a comparable, comprehensive process to select, train, and establish calibration of the assessors who score candidate submissions.
- Claim 5: The standard-setting procedures used for each TPA model are sufficiently comparable and rigorous to ensure that the respective passing standards for each model accurately and consistently identify candidates possessing the requisite knowledge, skills, and abilities (KSAs) required to effectively teach the content area(s) authorized by the credential.
- Claim 6: The model sponsor for each TPA model conducts statistical analyses to identify differential effects in relation to candidates' race, ethnicity, language, gender or disability. Any differences are documented, and processes implemented to eliminate sources of construct-irrelevant variance.
- Claim 7: For each TPA model, the score reports (candidate-level and program-level) provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used.
- Claim 8: The rubrics and score reports provide diagnostic information on candidates and on programs such that the strengths and weaknesses of each can be identified.

Seven distinct activities (studies) were designed to investigate these claims.[3] The TAC provided guidance on the design, implementation, and interpretation of results for the seven activities. An overview of each activity and its results is presented next.

[2] Claims 1 - 7 were identified in the HumRRO proposal and ensuing work plan. The claims were reviewed and approved by the Commission and TAC with minor edits. Claim 8 was added at the project kick-off meeting at the request of the Commission.
[3] See Table 8.1 in Chapter 8 (Summary) for a table mapping Claims to activities/studies, along with overall findings.

Evaluation and Comparison of Evidence across TPA Models for Adherence to Assessment Design Standards (Activity 1)

Activity 1 involved a comprehensive review and comparison of the documents and materials developed by each model sponsor, such as technical specification documents, item bias review reports, scorer training materials, and sample score reports. In Year 1 (2017–18) of the comparability investigation, the available technical documentation for FAST and CalTPA, which were both being field tested, was limited and sparse. On the other hand, the available technical documentation for edTPA—which did not require substantive revisions in light of the revised ADS and TPEs—was more robust in Year 1. In Year 2 (2018–19), additional and more detailed documentation became available for FAST and CalTPA. Moreover, additional detail and clarification on the available documentation was provided for all three models in Year 2. As a result, the average ratings for adherence to Standards increased from Year 1 to Year 2, particularly for FAST and CalTPA, which had comparatively lower ratings than edTPA in Year 1. With these improvements, the technical documentation indicates that all three TPA models mostly or fully adhere to the ADS. This provides support for Claim 1, which states, in part that, "*The TPA models are sufficiently comparable in their representation of the Commission's Assessment Design Standards*." We further determined that the models mostly or fully adhere to relevant standards from *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter referred to as the *Joint Standards*), which are generally accepted as industry-wide principles for test design and development.

Content Validity Comparability Analysis (Activity 2)

Activity 2 built upon Activity 1 by further investigating Claim 1. Specifically, Activity 2 investigated the part of Claim 1 which states that, "*The TPA models are sufficiently comparable in their assessment and weighting of the Commission-adopted TPEs*." Two independent panels of teacher preparation experts mapped the components of each TPA model to each TPE element in spring 2018. A subset of the panelists reconvened in summer 2019 to revisit and update the linkages/mappings from spring 2018. Overall, the findings for Activity 2 indicate that there are some differences in the emphasis and measurement of TPEs across the TPA models; however, there is more comparability than dissimilarity, particularly between FAST and CalTPA. This provides partial but not full support for the claim that the TPA models are sufficiently comparable in their assessment and weighting of the TPEs (Claim 1). The ADS state that each TPA task must be "substantively related to two or more major domains of the TPEs" and that "collectively, the tasks and rubrics in the assessment address key aspects of the six major domains of the TPEs" (ADS 1a). All models meet this Standard. However, the ADS do not specify which of the 45 TPE *elements*, nor how many of them, each model must measure. Thus, given that the ADS allow model developers considerable flexibility in deciding which TPE elements to assess, it is perhaps not surprising that a mapping of TPE elements to TPA components revealed some differences across the three TPA models.

Comparison of Stakeholder Input across TPA Models (Activity 3)

The primary purpose of Activity 3 was to investigate Claim 2: "*The guidance and supports (e.g., guide/manual/handbook and other resources) provided by model sponsors to candidates and teacher preparation faculty are sufficiently clear and detailed to ensure that the model is implemented as designed and intended*." To investigate this claim we gathered stakeholder perceptions through online surveys. Survey results indicate that the majority of candidates across all three models agree that they understand the requirements (e.g., directions, rubrics, evidence requirements) for their TPA model, although self-reported understanding of

requirements appeared to be strongest for FAST candidates. There was also consistently strong agreement across models that program coordinators have a clear understanding of their model's purpose and requirements, and that they felt well informed during the assessment process. Furthermore, the survey findings indicate that the TPA models are perceived as valid by both candidates and coordinators across all three models. These findings should help to ensure that the TPA models are implemented as designed and intended, and thereby lend support to Claim 2.

Scoring Review – Comparison of Scoring Rubrics, Score Reports, and Rater Training (Activity 4)

Activity 4 was an evaluation of the quality and comparability of scoring procedures and scoring-related materials, and it investigated Claims 3, 4, 7 and 8. HumRRO staff with scoring expertise conducted an extensive review of scoring materials, observed scorer training, and interviewed key scoring personnel from each model. A review of each model's scoring rubrics (Claim 3) reveals that they are sufficiently clear and detailed to ensure that trained scorers can accurately and consistently score candidate submissions, although FAST, whose rubrics contain aspects of analytic and holistic scoring, may benefit from providing additional guidance to scorers on how to collapse over indicator level ratings to arrive at an overall rating for each rubric. Next, a review of each model's scorer training (Claim 4) showed that scorer trainings for all TPA models address key aspects of the ADS and *Joint Standards* related to scorer training, although edTPA and CalTPA have stronger procedures to ensure that returning scorers are calibrated and that scorers remain calibrated throughout the scoring window. Next, a review of score reports for information on candidate outcomes and intended score use (Claim 7) revealed that score reports for CalTPA and edTPA are similar in that they (a) provide information on the candidate's passing status (although for edTPA this is through a weblink included on score reports), and (b) include guidance that scores should be used to compare candidates' performance (knowledge and skills) to the requirements set by the Commission/their state. This information is not included on FAST score reports, but it is included in the FAST Manual. None of the models included guidance in their score reports that scores should be used in conjunction with other measures to determine a candidate's readiness for beginning teaching, although all models included this information in other supporting materials. Finally, score reports for all models are diagnostic (Claim 8) in the sense that they provide rubric-level scores (for both candidates and programs). However, only CalTPA score reports explicitly state that rubric-level scores "may help you identify your relative strengths and areas of improvement," while the FAST and edTPA models provide similar guidance in other supporting materials.

Comparison of Standard Setting across TPA Models (Activity 5)

The purpose of Activity 5 was to investigate Claim 5: "*The standard-setting procedures used for each TPA model are sufficiently comparable and rigorous to ensure that the respective passing standards for each model accurately and consistently identify candidates possessing the requisite KSAs required to effectively teach the content area(s) authorized by the credential*." We used direct observation and documentation provided by the model sponsors to evaluate standard-setting procedures. After review of the standard-setting evidence for all three TPA models, we concluded that edTPA and CalTPA procedures are sufficiently comparable and rigorous to ensure that their passing standards accurately and consistently identify candidates possessing the requisite KSAs required to effectively teach the content area(s) authorized by the credential. Both models (a) appropriately considered the judgements of a suitable set of educators regarding an acceptable passing standard using a similar implementation of the briefing book standard setting method, (b) utilized performance data (i.e., impact data) and

candidate score profiles to inform judgements, (c) documented their process at a similarly deep and appropriate level, and (d) framed the need of each panelist to create a definition of KSAs associated with minimally qualified candidates in a similar manner. The procedures used by FAST were not as rigorous as those used by edTPA and CalTPA. The FAST model used a nontraditional standard setting method whereby teacher preparation staff at Fresno State reviewed the Level 1 ("Does Not Meet Expectations") and Level 2 ("Meets Expectations") rubric descriptors to ensure that the Level 2 descriptors adequately described the KSAs of a just-sufficiently-qualified beginning teacher. As such the "cut score" was identified as a Level 2 rating on all 10 rubrics.

Statistical Analysis and Comparison of Score Data across TPA Models (Activity 6)

Activity 6 was conducted as an independent investigation of the veracity of Claim 6, which states that "*The model sponsor for each TPA model conducts statistical analyses to identify differential effects in relation to candidates' race, ethnicity, language, gender or disability*." Claim 6 stems from ADS 1(k) which states, "*The model sponsor completes initial and periodic basic psychometric analyses to identify pedagogical assessment tasks and/or scoring rubrics that show differential effects in relation to candidates' race, ethnicity, language, gender or disability*." To investigate Claim 6, we compared pass rates and total scores across models by race and gender. We focused on race and gender because these data were available from all three models. We found no evidence to suggest substantive differences in pass rates for males and females within TPA models. Moreover, the pattern of pass rates for males and females was comparable across models. In addition, when we examined differences in mean total scores the magnitude of the differences between males and females were similarly small for all three models. These findings support the claim that there are no differential effects in relation to candidates' gender (Claim 6). The findings examining differences across racial groups were more complex. First, the racial demographics of the FAST population differ significantly from the racial demographics of the edTPA and CalTPA populations; the majority of the FAST candidates are Hispanic rather than White. Thus, FAST was not included in the analyses comparing differences between models on pass rates by race. The findings indicate that the pass rates for the various race categories were similar both within and across models for edTPA and CalTPA, thereby lending support to the claim that there are no differential effects in relation to candidates' race (Claim 6). Comparisons of mean total scores showed no notable differences among racial groups for any of the models except that White candidates tended to have higher mean total scores on FAST than Hispanics candidates, but all candidates passed the FAST TPA and, thus, the differences in mean total scores did not translate into differences in pass rates. These results are based on the multiple-subject credential only and should be revisited as more data (for multiple subject and for other credential areas) becomes available.

Comparison of TPA Models to a Common Criterion (Activity 7)

The final activity (Activity 7) represented an innovative and informative method for investigating the ultimate question of comparability across TPA models. We used the results from Activity 2—the content validity investigation—to identify a list of TPE elements that are assessed in substantively the same way across the TPA models and for which all the models measure the full depth and breadth of those TPE elements. We then developed a "Common Rubric" to measure those common TPE elements. Trained and calibrated assessors scored a representative sample of candidate submissions (for the multiple subject credential) from each model using this Common Rubric. We then conducted comparability analyses across TPA models using the scores on the Common Rubric as a referent. The findings indicate that scores on the Common Rubric were moderately strongly to strongly correlated with the scores from

each model's rubric. This suggests that, despite the unique components and rubrics for each TPA model, all three models are measuring a highly related construct of teaching performance (based on the subset of TPEs that could be reliably compared). To further explore the comparability of the scores, we regressed the Common Rubric scores onto the Model Rubric scores to identify a predicted cut score on the Common Rubric for each model. We computed the 95% confidence interval around each predicted cut score. The findings indicate that the models' confidence interval ranges of cut scores overlap for each model, suggesting that the three models would comparably classify candidates as passing or failing. In other words, regardless of which teaching performance assessment a candidate completes, his/her performance is likely to be consistently classified as passing or failing by all three models (again, based on the subset of TPEs that could be reliably compared). A classification consistency analysis also showed that the great majority of portfolios were consistently scored as "passing" on both rubrics or consistently scored as "failing" on both rubrics. Collectively, the findings from these analyses support that the pass/fail outcomes from each model are comparable when compared to a common criterion measure, although it is important to note that this conclusion is based on a small subset of TPEs on which all TPAs could be reliably compared.

Summary and Recommendations by Claim

A summary of the evidence for each claim is presented next along with some recommendations that may further strengthen support for the claims.

Claim 1: The TPA models are sufficiently comparable in their representation of the Commission's Assessment Design Standards (ADS) and in their assessment and weighting of the Commission-adopted Teaching Performance Expectations (TPEs).

The findings from **Activity 1** indicate that all three models mostly or fully adhere to the ADS. With regard to the assessment and weighting of TPEs, the findings from **Activity 2** indicate that each task/cycle for each model substantively assesses two or more TPE domains and that the tasks/cycles and rubrics for each model collectively address key aspects of the six TPE domains; these are requirements of ADS 1(a) and all models adhere to these requirements. We found that TPE 3 is the domain assessed most thoroughly by all three TPAs and TPE 6 is the domain assessed least thoroughly by all three TPAs. We did find that all models assessed the full depth and breadth of TPE element 6.1, but none of the models assessed key aspects of any of the other six elements comprising TPE 6.[4] The teacher preparation experts that participated in Activity 2 commented that the KSAs described in TPE 6 are difficult to measure via a performance assessment. Thus, it is important for the programs to ensure that the breadth of TPE 6 is being addressed through other means besides the performance assessment. The findings from Activity 2 also indicate that edTPA assesses key aspects of several TPE elements within TPE domain 2 (Creating and Maintaining Effective Environments for Student Learning) and TPE domain 4 (Planning Instruction and Designing Learning Experiences for All Students) but does not cover the full depth and breadth of any of the TPE elements within these two domains. Again, ADS 1(a) requires that each model "address key aspects" of each TPE domain, and, thus, edTPA complies with this requirement, but there may be opportunity for edTPA to further strengthen the assessment of the TPE elements within TPE domains 2 and 4, based on

[4] See page 59 in the Method section of Chapter 2 for an explanation of how "full depth and breadth" and "key aspects" were defined for this activity.

the findings from Activity 2.[5] In summary, Activity 2 revealed comparability across models in their assessment of TPE *domains*, but some differences with regard to each model's assessment and weighting of TPE *elements* within each domain. Finally, survey findings from **Activity 3** indicate that the majority of candidates and program coordinators perceive their TPA model as valid. Collectively, these findings lend support to Claim 1.

Claim 2: The guidance and supports (e.g., guide/manual/handbook and other resources) provided by model sponsors to candidates and teacher preparation faculty are sufficiently clear and detailed to ensure that the model is implemented as designed and intended.

The evidence for this claim comes from online surveys administered to candidates and program coordinators (**Activity 3**). The results from the surveys indicate that the majority of candidates and coordinators agree that they understand the requirements (e.g., directions, rubrics, evidence requirements) for their TPA model and that the resources and supports provided to them by their model sponsor are helpful. Only FAST candidates reported difficulty with one resource—the online system for uploading their submissions. We recommend that the FAST model sponsor investigate ways to improve the online submission system for uploading candidate portfolios. The survey results also revealed that the TPA models are perceived as valid by both candidates and coordinators across all three models. These findings should help to ensure that the TPA models are implemented as designed and intended, and thereby lend support to Claim 2.

Claim 3: The scoring rubrics for each TPA model are sufficiently clear and detailed to ensure that trained scorers can accurately and consistently score candidate submissions.

After review of the available information for FAST's, edTPA's, and CalTPA's rubrics (**Activity 4**), we found that, overall, the rubrics are sufficiently clear and detailed to ensure that trained raters can accurately and consistently score candidate submissions, and that all the TPAs mostly or fully adhere to the ADS and *Joint Standards* that are relevant to rubrics. The format and structure of the edTPA and CalTPA rubrics are similar; both are analytic with five scoring levels labeled Level 1 through Level 5. FAST uses four score levels with labels that range from "Does Not Meet Expectations" to "Exceeds Expectations." Each of FAST's 10 rubrics contain 2-3 indicators. We recommend that FAST develop written guidance for assessors on how to weight indicators within each rubric. A short guideline for determining how to weigh the importance of individual indicators for overall rubric ratings could help to further strengthen the reliability of FAST scores. We also recommend that edTPA consider making the linkage between TPE elements and edTPA rubrics and tasks readily available to candidates and programs (as do FAST and CalTPA), perhaps via a supplemental linkage document; this may also prove useful for improving assessors' understanding of TPE elements. Finally, we recommend that all TPA models ensure all levels of their rating scale are presented to assessors at training. Exemplars at the extremes of the scales (e.g., Level 1 and Levels 4 and 5) were noticeably underrepresented at observed assessor training sessions and in assessor training materials.

[5] We note that edTPA is a nationally validated assessment program used in 44 states. edTPA was initially approved for use in California in 2014 having illustrated alignment to the TPEs. In 2015 the Commission adopted revised Assessment Design Standards and in 2016 the Commission adopted revised TPEs. The edTPA was again approved by the Commission in 2018, having demonstrated alignment to the TPEs.

Claim 4: For each TPA model, there is a comparable, comprehensive process to select, train, and establish calibration of the assessors who score candidate submissions.

A review of each model's scorer training (**Activity 4**) showed that scorer trainings for all TPA models address key aspects of the ADS and *Joint Standards* related to scorer training, although edTPA and CalTPA have stronger procedures to ensure that returning scorers are calibrated and that scorers remain calibrated throughout the scoring window. Returning scorers for FAST were not required to re-calibrate in 2018–19, although they did attend training sessions to discuss the revisions that were made to the rubrics following field test. We recommend FAST require all its scorers, including returning scorers, to re-establish calibration on a qualifying portfolio, especially when revisions, even minor ones, are made to rubrics and/or tasks. We also recommend that FAST incorporate a calibration exercise near the middle of each scoring window to ensure all scorers are still scoring consistently and are calibrated. This exercise, even if brief, would help to identify and correct scorer drift, a common phenomenon in extensive scoring activities.

Claim 5: The standard-setting procedures used for each TPA model are sufficiently comparable and rigorous to ensure that the respective passing standards for each model accurately and consistently identify candidates possessing the requisite KSAs required to effectively teach the content area(s) authorized by the credential.

It appears that edTPA and CalTPA use standard setting procedures (Briefing Book method) that are sufficiently comparable and rigorous to ensure that their passing standards accurately and consistently identify candidates possessing the requisite KSAs required to effectively teach the content area(s) authorized by the credential (**Activity 5**). The procedures used by FAST were not comparable to, nor as rigorous as, those used by edTPA and CalTPA. While an excellent review of the clarity and appropriateness of the rubrics, future FAST standard-setting activities should consider including performance data (i.e., impact data), actual candidate submissions representing a variety of performance levels, and consideration of a compensatory scoring model in order to make its standard setting more rigorous and comparable to edTPA and CalTPA.

Claim 6: The model sponsor for each TPA model conducts statistical analyses to identify differential effects in relation to candidates' race, ethnicity, language, gender or disability. Any differences are documented, and processes implemented to eliminate sources of construct-irrelevant variance.

Overall, the findings from **Activity 6** support the claim that there are no differential effects in pass rates in relation to candidates' gender or race. We recommend that model sponsors include additional demographic information in their score files so that ADS 1(k), which states that, "*The model sponsor completes initial and periodic basic psychometric analyses to identify pedagogical assessment tasks and/or scoring rubrics that show differential effects in relation to candidates' race, ethnicity, **language**, gender or **disability***" [emphasis added] can be more fully investigated.

Claim 7: For each TPA model, the score reports (candidate-level and program-level) provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used.

All models provide rubric level scores on candidate score reports and program score reports (**Activity 4**). CalTPA and FAST also include total scores and information on passing status

(although for edTPA passing status is accessed through a weblink included on the score report). FAST does not include a total score on its score reports, but all candidates know that they must obtain at least a '2' on all 10 rubrics in order to pass FAST. All three models may want to consider including additional guidance on their score reports regarding appropriate score use. For example, none of the models include guidance on their score reports that scores should be used in conjunction with other measures to determine a candidate's preparedness for beginning teaching. All three models include this information in their reference materials, but not directly on score reports. Ideally, a score report would be a self-contained document that does not require review of reference or supporting materials for information on how score information should be used.

Claim 8: The rubrics and score reports provide diagnostic information on candidates and on programs such that the strengths and weaknesses of each can be identified.

Score reports (**Activity** 4) for all models are diagnostic in the sense that they provide rubric level scores (for both candidates and programs), but only CalTPA score reports specifically state that rubric level scores "may help you identify your relative strengths and areas of improvement." The FAST and edTPA models do provide similar guidance in other supporting materials. The FAST and edTPA models may want to consider including similar guidance in their score reports. We also recommend that models convey that rubric scores and overall scores be used in conjunction with other information to make determinations about a candidate's readiness for beginning teaching.

Practical Implications

The primary practical implication of this investigation is that it provides empirical evidence to support the Commission's decision to approve multiple TPA models as a credentialing requirement for beginning teachers. Again, this is not to say that the models are equal, but rather that all models are likely to equitably identify teacher candidates who are "ready"—that is, possess the KSAs required for beginning teaching. The findings from this investigation do point out some potential threats to the comparability of the TPAs, which the model sponsors are encouraged to address. Doing so will further strengthen model comparability, as well as the quality and rigor of the TPA model. If the Commission is concerned about differences across the models in the representation of the TPE *elements* assessed, then to further strengthen model comparability the Commission might consider providing the model sponsors with guidance at the level of TPE *elements*, rather than just TPE domains. This could be done through a modification to the *Assessment Design Standards*. This investigation shows that the ADS have provided a strong blueprint for the models to follow and that the model sponsors are closely adhering to the ADS. This suggests that any changes the Commission might make to the ADS are likely to be enacted by the model sponsors.

Future Research

This body of research demonstrates a comprehensive investigation of TPA model comparability. Nonetheless, additional research is recommended to further support the validity argument for model comparability. Validity arguments are not static, rather they are dynamic and are strongest when supported by ongoing research to support continuous improvement. Suggestions for future research include an expansion or elaboration upon the studies conducted herein as well as new avenues of research. Some of the ways the studies described in this report could be expanded or elaborated upon are outlined below.

- Conduct another content validity investigation (Activity 2) but expand upon it by having teacher preparation experts identify which aspects of each TPE *element* are assessed by each model. In the current effort, a strong evidence linkage indicated that the model assessed the full depth and breadth (i.e., all aspects) of the TPE element. Thus, these were the TPE elements included in the Common Rubric in Activity 7. A moderate evidence linkage indicated that the model assessed key aspects of the TPE element, but not the full depth and breadth. Because we wanted to ensure that all models would be compared on a level playing field in Activity 7, only TPE elements for which there was "strong evidence" across all three models were included in the common rubric. However, if the key aspects of the TPE elements that received "moderate evidence" ratings were identified, then additional teaching performance expectations that are common across models could be identified.
- The above bullet point could be further extended by creating a new common rubric that more fully addresses the construct space, and then updating Activity 7 using this more robust common rubric.
- Activity 6 (investigation of score patterns for subgroups) could be conducted for other credential areas beyond the multiple subject credential. Also, Activity 6 could be expanded upon by investigating subgroup differences in score patterns for language and disability, assuming the models capture this demographic information in their score data. If multiple years of data were combined, then this would help to circumvent concerns regarding small samples.
- When/if notable changes are made to a model(s), any number of the seven activities could be repeated to evaluate the improved model.

There are also new avenues of research that could supplement this existing body of research. Some new areas of research might include:

- A longitudinal predictive validity study in which candidates' scores on their TPA are correlated with a measure of their teaching performance (e.g., their performance evaluation from their first year of teaching). Such a study would address an important gap in the validity argument for the TPA models—i.e., it would provide empirical evidence that the models are indeed predictive of the KSAs necessary for beginning teachers.
- A convergent validity study in which candidates' scores on the TPA are correlated with another assessment of teaching. Another assessment to potentially explore is the California Basic Educational Skills Test (CBEST), which measures candidates' content knowledge in the areas of reading, mathematics and writing, and for which scores are readily available to the Commission. One might expect that candidates' scores on their single subject mathematics portfolio, for example, may correlate more strongly with their scores on the mathematics portion of CBEST (convergent evidence) than with their scores on the reading portion of CBEST (discriminant evidence)—i.e., teachers must know the content areas they teach. Such information would help to support the construct validity evidence for the TPAs and for the CBEST alike.

The research listed above would not only further support the validity argument for model comparability but would also further strengthen the validity evidence for any given model.

Final Conclusion

As with all research studies, these seven activities are not without their limitations, and those are described within the body of the report. Nonetheless, this investigation paints a rich picture of comparability. Certainly, there are differences across the TPA models. In many cases those differences do not pose threats to the veracity of the claims and the differences are in line with the Commission's expectations—as evidenced by the fact that the Commission's *Assessment Design Standards* allow for flexibility in how each model assesses the TPEs. However, some of the identified differences may pose threats to the veracity of the claims and, ultimately, to the equitable identification of "TPE-ready professionals." In this regard, this report should serve a formative purpose for the model sponsors so that they can address potential threats to model comparability.

In conclusion, the Commission should be commended for undertaking a comprehensive investigation of the comparability of the TPA models. Not only does this investigation bolster support for the claim that the TPA models are comparable, it also strengthens the validity evidence for each of the models. As such, the Commission can be assured that there is compelling validity evidence to support each of the models they have approved. As one of the TAC members commented, this investigation may serve as a useful roadmap for other states and/or credentialing organizations that are considering approving multiple performance assessments for credentialing decisions.

An Investigation of the Comparability of Commission-Approved Teaching Performance Assessment Models: Final Report

Table of Contents

Table of Contents (Continued)

Table of Contents (Continued)

List of Tables

Table of Contents (Continued)

List of Tables

Table of Contents (Continued)

List of Tables

List of Figures

Table of Contents (Continued)

List of Figures

An Investigation of the Comparability of Commission-Approved Teaching Performance Assessment Models: Final Report

Introduction

California's Commission on Teacher Credentialing (Commission) requires all programs of preliminary multiple and single subject teacher preparation to use a Commission-approved Teaching Performance Assessment (TPA) as one of the program completion requirements for prospective teacher candidates. In conformance with applicable California statute, multiple TPA models are allowed across the state as a program completion requirement for prospective teacher candidates. There are three TPA models approved by the Commission. They are:

- the FAST (Fresno Assessment of Student Teachers), owned and operated by Fresno State; and
- the CalTPA (California Teaching Performance Assessment), originally developed by Educational Testing Service (ETS) and owned by the Commission, revised by a Design Team with an operational contractor of the Evaluation Systems group of Pearson; and
- the edTPA, owned by Stanford University, with an operational contractor of the Evaluation Systems group of Pearson.

Each TPA model must meet the Commission's *Assessment Design Standards* (adopted December 2015)[6] and measure the Commission-adopted *Teaching Performance Expectations* (adopted June 2016).[7] The *Assessment Design Standards* (ADS) describe the design requirements for all TPA models, as set forth by the Commission. The *Teaching Performance Expectations* (TPEs) describe the performance standards for beginning teachers. There are six TPE "domains" (e.g., TPE domain 1: *Engaging and Supporting all Students in Learning*) and each domain includes six to eight descriptors, referred to as "elements," which describe the knowledge, skills, and abilities (KSAs) required for beginning teachers.

Although each TPA model is Commission-approved, they differ in several important ways (e.g., design of candidate tasks, scoring rubrics, teaching performance elements measured by tasks). These inter-model differences raise questions regarding the comparability of results obtained by teacher candidates completing the various TPAs. Consequently, the Commission contracted with the Human Resources Research Organization (HumRRO) to conduct an external, independent investigation of the comparability of the three TPA models. The investigation occurred between June 2017 and December 2019.

Given the Commission's adoption of the ADS in December 2015 and the TPEs in June 2016, each of the models underwent revisions. The FAST and CalTPA models have required more extensive revisions than edTPA, constituting the need for pilot testing in 2016–17 and field testing in 2017–18. Thus, Year 1 (2017–18) of the HumRRO investigation focused on the models as they were being revised (i.e., field tested for CalTPA and FAST) and Year 2 (2018–19) focused on the operational models (i.e., post-field test).

[6] https://www.ctc.ca.gov/docs/default-source/educator-prep/tpa-files/tpa-assessment-design-standards.pdf
[7] https://www.ctc.ca.gov/docs/default-source/educator-prep/standards/adopted-tpes-2016.pdf

Technical Approach

Our goal was to investigate the comparability of the three TPAs. Numerous techniques can be used to compare assessments. For example, an equating study is a rigorous technique to derive a function to map a score from one test onto the scale of another test, so that any given score has the same meaning regardless of which test was administered. However, equating requires that the tests being compared have either some common items, or that some people take both tests. Unfortunately, the TPA models do not share items nor examinees. Thus, the objective of this investigation was to compare the three TPAs on key aspects of test design, implementation, scoring, and reporting to create a body of evidence and thereby triangulate—that is, capture from different angles—whether the models are indeed comparable. The goal was to accumulate as much evidence as possible (i.e., a "body of evidence") to evaluate the comparability of the three TPA models.

This investigation adopted a "Theory of Action approach" (Kane, 2006; 2013) to identify the claims that need to be substantiated to "assure that the Commission-approved TPA models are *sufficiently comparable* [emphasis added] that they are equitably assessing candidates working toward a California preliminary multiple or single subject teaching credential" (Request for Proposal, p. 5). This investigation was guided by a technical advisory committee (TAC) comprised of model sponsors and independent assessment experts. During the first TAC meeting the attendees engaged in a discussion of the meaning of "sufficiently comparable." This discussion resulted in the following guidance: "comparable does not mean that the models are equal in *how* they measure the KSAs required by the TPEs, but that all models equitably identify TPE-ready professionals." To assure that this ultimate objective is attained the following claims must be substantiated:

- Claim 1: The TPA models are sufficiently comparable in their representation of the Commission's *Assessment Design Standards* (ADS) and in their assessment and weighting of the Commission-adopted *Teaching Performance Expectations* (TPEs).
- Claim 2: The guidance and supports (e.g., guide/manual/handbook and other resources) provided by model sponsors[8] to candidates and teacher preparation faculty are sufficiently clear and detailed to ensure that the model is implemented as designed and intended.
- Claim 3: The scoring rubrics for each TPA model are sufficiently clear and detailed to ensure that trained scorers can accurately and consistently score candidate submissions.
- Claim 4: For each TPA model, there is a comparable, comprehensive process to select, train, and establish calibration of the assessors who score candidate submissions.
- Claim 5: The standard-setting procedures used for each TPA model are sufficiently comparable and rigorous to ensure that the respective passing standards for each model accurately and consistently identify candidates possessing the requisite knowledge, skills, and abilities required to effectively teach the content area(s) authorized by the credential.

[8] Per the *Assessment Design Standards,* "model sponsor" refers to the entity that represents the assessment and is responsible to programs using that model and to the Commission.

- Claim 6: The model sponsor for each TPA model conducts statistical analyses to identify differential effects in relation to candidates' race, ethnicity, language, gender or disability. Any differences are documented, and processes implemented to eliminate sources of construct-irrelevant variance.

- Claim 7: For each TPA model, the score reports (candidate-level and program-level) provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used.

- Claim 8: The rubrics and score reports provide diagnostic information on candidates and on programs such that the strengths and weaknesses of each can be identified. [9]

Seven different activities (studies) were designed to investigate these claims. The aforementioned TAC provided guidance on the design, implementation, and interpretation of results for these seven activities.

Purpose of the Current Report

Over the course of this 2.5-year investigation, HumRRO submitted intermittent progress reports, a Year 1 preliminary report (Sinclair & Thacker, 2018), and this culminating document, the Year 2 final report. In the chapters that follow, we present the findings from the following seven activities that were completed for this comparability investigation:

- Activity 1: Evaluation and Comparison of Evidence across TPA Models for Adherence to *Assessment Design Standards* (Chapter 1)

- Activity 2: Content Validity Comparability Analysis (Chapter 2)

- Activity 3: Comparison of Stakeholder Input across TPA Models (Chapter 3)

- Activity 4: Scoring Review—Comparison of Scoring Rubrics, Scorer Training, and Score Reports across TPA Models (Chapter 4)

- Activity 5: Comparison of Standard Setting across TPA Models (Chapter 5)

- Activity 6: Statistical Analysis and Comparison of Score Data across TPA Models (Chapter 6)

- Activity 7: Comparison of TPA Models to a Common Criterion Measure (Chapter 7)

[9] Claim 8 was added to the list of claims as a result of discussion with the Commission at the project kick-off meeting on June 22, 2017.

Chapter 1: Evaluation and Comparison of Evidence across TPA Models for Adherence to Assessment Design Standards (Activity 1)

Emily Dickinson, Andrea Sinclair & Justin Paulsen

Introduction

Activity 1 is as an overarching investigation, via a documentation review, of the eight claims identified in the Introduction. Activity 1 most directly addresses Claim 1, "*The TPA models are sufficiently comparable in their representation of the Commission's Assessment Design Standards . . .*" Activity 1 involved a comprehensive review and comparison of the documents and materials developed by each model sponsor. This evidence was reviewed, and evaluations were made regarding the strength of evidence for adherence to the *Assessment Design Standards* (ADS) and *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter referred to as the *Joint Standards*). The latter evaluation was conducted to ensure that the documentation and materials for each TPA model also satisfy industry-wide principles for test design and development.

This same activity was conducted in Year 1 (2017-2018) of the Comparability Study. Because the models were undergoing revisions in 2017-2018, particularly for FAST and CalTPA, the findings presented in the Year 1 report (Sinclair & Thacker, 2018) were presented as preliminary. The revised models became operational in 2018-2019.[10] Thus, Activity 1 was updated in Year 2 of the Comparability Study (2018-2019) and the findings presented herein represent the final evaluation for Activity 1 of the Comparability Study.

Method

Year 1 Process (2017-2018)

Using the ADS as the guiding framework, the HumRRO project team created a list of documentation and materials to request from each of the model sponsors to allow us to make determinations about the strength of evidence for adherence to the ADS. Each model sponsor was emailed an Excel spreadsheet that included the list of requested materials. They were asked to use the spreadsheet to identify the specific documentation that addressed each element of our request; a column was provided next to the request in which the model sponsor was asked to enter the names of the documents (or weblinks where the information could be found) that addressed each request. The spreadsheet included a column for the model sponsor to enter any additional comments or explanations about the documentation they provided. A column was also provided to allow for explanations of why particular documentation may not be available at this time. The model sponsors were given the option to upload the requested materials to HumRRO's secure ftp site or to use their own secure site for posting any secure materials.

One key project staff member was assigned to review the documentation and materials provided by each model sponsor. That person then mapped the documentation and materials onto each of the ADS to which the materials were relevant. This mapping was then shared with each of the model sponsors to ensure that (a) we had accurately captured the information they provided and (b) to give the model sponsors the opportunity to provide any additional information that may be relevant to addressing the ADS. After receiving feedback from the

[10] edTPA underwent minor revisions in 2017–18. Thus, unlike CalTPA and FAST, edTPA did not require field testing in 2017–18.

model sponsors and any additional information that was provided with that feedback, the key staff member assigned to each TPA model proceeded to review the materials and make evaluations about the strength of evidence for each ADS. This same process was repeated for the relevant test design and development Standards from the *Joint Standards*.[11]

Additional chapters of this report (i.e., the Comparability Study activities), also address the ADS and the *Joint Standards.* For example, Chapter 4 is an in-depth review of scoring. To minimize redundancies across chapters, the Standards that are substantively investigated in other chapters of this report are not included here in Chapter 1.

Year 2 Process (2018-2019)

The Activity 1 chapter from the Year 1 report was shared with the model sponsors. Model representatives were asked to review the evaluations about the strength of evidence for each ADS and *Joint Standard* (along with the supporting rationale for the rating) and provide any updated and/or additional documentation/evidence for Year 2.

Ratings

Evaluations in both Year 1 and Year 2 were made using the rating scale shown in Table 1.1. All ratings were supported by a corresponding rationale. All key staff members responsible for evaluating each model were trained and calibrated on the rating scale. The HumRRO project director then conducted a cross-check on all ratings and rationales; any discrepancies were discussed with the rater to reach a consensus.

Table 1.1. Rating Scale for Strength of Evidence

Rating Level	Description of Rating Level
1	No evidence of the Standard/element found in the documentation provided.
2	Little evidence of the Standard/element found in the documentation; less than half of the Standard/element covered in the documentation and/or evidence of key aspects of the Standard/element could not be found.
3	Some evidence of the Standard/element found in the documentation; approximately half of the Standard/element covered in the documentation including some key aspects of the Standard/element.
4	Evidence in the documentation mostly covers the Standard/element; more than half of the Standard/element covered in the documentation, including key aspects of the Standard/element.
5	Evidence in the documentation fully covers all aspects of the Standard/element.

[11] We capitalize “Standard” throughout this report when referring to a standard specified by the ADS or the *Joint Standards*, as opposed to a standard that is a generally accepted expectation in the industry.

Results

Appendix 1.A presents the information we requested for each TPA model and the subsequent documentation and materials provided by the model sponsors.[12] The model sponsors provided a variety of materials and formats such as, PowerPoint presentations, technical reports, and weblinks to information and materials posted on the internet.

Evaluation of Strength of Evidence for Assessment Design Standards

Next, we present the results of the numeric ratings assigned to each of the 28 ADS using the rating scale presented in Table 1.1. Each table describes one model and includes the (a) ADS in the left column, (b) rating on the strength of evidence for the Standard in the middle column, and (c) rationale for the rating in the right column.

FAST. Table 1.2 presents the ratings for FAST on each ADS. In Year 1 (2017–18), much of the technical documentation requested for FAST was not yet available. Thus, in Year 1, only 7 of the 28 ADS could be rated. The average rating for those seven ADS in Year 1 was 4.29 on the 5-point rating scale. In Year 2 (2018–19), there was considerably more technical documentation available for FAST; all but one ADS was rated. All Standards were rated a '4' (Evidence in the documentation mostly covers all aspects of the Standard/element) or '5' (Evidence in the documentation fully covers all aspects of the Standard/element), with the majority receiving a rating of a '5' on the 5-point rating scale. The average rating across the ADS for Year 2 was 4.83, which is a notable increase from Year 1.

[12] Appendices for this report are in Volume II: Appendices.

Table 1.2. Ratings on the Assessment Design Standards for FAST

Assessment Design Standard	FAST Rating	Rationale for FAST Rating
1(a) See Chapter 4		
1(b) The TPA model sponsor must include a focus on content-specific pedagogy within the design of the TPA tasks and scoring scales to assess the candidate's ability to effectively teach the content area(s) authorized by the credential.	5	Per the FAST Response to the Assessment Design Standards document (p. 4), for Multiple Subject candidates, the SVP includes a focus on content-specific pedagogy related to literacy within the context of an integrated unit. For Single Subject candidates, the focus for both tasks is their specific content area. As part of their self-evaluation in the SVP, candidates select a video clip from their lesson demonstrating the use of subject-specific pedagogy and provide a justification for their choice. In the TSP, candidates plan, implement and reflect on a standards-based unit of study with a focus on disciplinary literacy. They must include California ELA/Literacy and/or California ELD standards for the literacy component of the unit. For multiple subject candidates, the focus must be an integrated unit using ELA standards as a key component.
1(c) See Chapter 4		
1(d) The model sponsor must include within the design of the TPA candidate tasks a focus on addressing the teaching of English learners, all underserved education groups or groups that need to be served differently, and students with special needs in the general education classroom to adequately assess the candidate's ability to effectively teach all students.	5	The SVP task requires candidates to address appropriate English Language Development (ELD) standards. The SVP also requires candidates to make use of data on students, and candidates score highest when they take into consideration information that is specific to individuals or student subgroups while identifying instructional choices. The FAST Tasks Matrix links the TSP task to TPEs related to "developmentally and ability-appropriate instructional strategies" and "supportive learning environment for students' first and/or second language acquisition." The TSP scoring rubrics are designed to reward candidates who tailor instruction to meet the needs of students with a wide range of English proficiency levels and/or identified special needs. Additional supporting rationale for this standard is provided on pgs. 7-9 of the FAST Response to the Assessment Design Standards document.
1(e) For Multiple Subject candidates, the model sponsor must include assessments of the core content areas of at least Literacy and Mathematics. Programs use local program performance assessments for History-Social Science and Science if not already included as part of the TPA.	5	The SVP task requires multiple subject candidates to conduct a Mathematics Lesson. The TSP task requires multiple subject candidates to focus on an integrated literacy unit. History/Social Science and Science are assessed within a *Disciplinary Literacies and Integrated Curriculum* course and a *Science Instruction and Applied Technology* course, respectively. Additional supporting rationale for this standard is provided on pg. 9 of the FAST Response to the Assessment Design Standards document.

(continued)

Table 1.2. (Continued)

Assessment Design Standard	FAST Rating	Rationale for FAST Rating
1(f) The model sponsor must include a focus on classroom teaching performance within the TPA, including a video of the candidate's classroom teaching performance with candidate commentary describing the lesson plan and rationale for teaching decisions shown and evidence of the effect of that teaching on student learning.	5	The SVP task requires candidates to plan, teach, and evaluate a 20-45-minute lesson that is observed by the University Coach and videotaped. The video must be submitted, along with the lesson plan and a reflection that is based on the candidate's review of the video. This reflection includes responses to questions related to the effectiveness of the lesson, along with evidence-based support. The TSP task includes a lesson plan, and a reflection and self-evaluation component that requires candidates to identify evidence of effective instruction and student learning. Additional supporting rationale for this standard is provided on pg. 10 of the FAST Response to the Assessment Design Standards document.
1(g) See Chapter 4		
1(h) See Chapter 4		
1(i) The model sponsor provides a clear statement acknowledging the intended uses of the assessment. The statement demonstrates the model sponsor's clear understanding of the implications of the assessment for candidates, preparation programs, the public schools, and TK-12 students. The statement includes appropriate cautions about additional or alternative uses for which the assessment is not valid. All elements of assessment design and development are consistent with the intended uses of the assessment for determining the pedagogical competence of candidates for Preliminary Teaching Credentials in California and as information useful for determining program quality and effectiveness.	4	The FASTv2.0 Complete Manual for teacher candidates contains an Intended Use Policy (p.41) that states that FAST is designed to provide (a) evidence of the pedagogical competence of Multiple and Single Subject Credential Candidates at Fresno State, as measured by the TPEs and (b) information useful for determining program quality and effectiveness. The manual also informs candidates of the minimum scores they must obtain on the FAST sections to receive course credit and describes the next steps for candidates who fail to achieve these minimum scores. The manual also states that FAST is just one of the requirements that teacher candidates must complete in order to earn their credential. The FAST Response to the Assessment Design Standards notes that FAST has not been released for use by other institutions and may only be used as designed (p. 13). There is no statement in the FAST documentation regarding the implications of FAST for public schools and TK-12 students.

(continued)

Table 1.2. (Continued)

Assessment Design Standard	FAST Rating	Rationale for FAST Rating
1(j) The model 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.	5	According to the FAST Response to the Assessment Design Standards (p.13), "prior to each field test, tasks and rubrics were reviewed for cultural sensitivity and for the use of academic language that might interfere with fairness for candidates from diverse backgrounds. Faculty representatives with expertise in linguistic and cultural sensitivity formally and critically reviewed all components of each task. The group of faculty content reviewers was composed of one Southeast Asian, one African American, one Hispanic, and one Caucasian. Tasks and rubrics were also reviewed by the FAST Transition Team, which includes experts in cultural and language foundations, students with special needs, educational psychology, and educational technology. In addition, two classes of graduate students from diverse backgrounds who had recently completed the teaching credential program at Fresno State reviewed the tasks and rubrics. Except for recommendations for minor changes in wording and formatting, all reviewers found the tasks to be culturally sensitive, fair, and appropriate."
1(k) The model sponsor completes initial and periodic basic psychometric analyses to identify pedagogical assessment tasks and/or scoring rubrics that show differential effects in relation to candidates' race, ethnicity, language, gender or disability. When group pass-rate differences are found, the model sponsor investigates the potential sources of differential performance and seeks to eliminate construct-irrelevant sources of variance.	5	Psychometric analyses were completed after field test to identify any differential effects in relation to candidates' ethnic group and gender. Analyses of pass rates, scoring reliability, and subgroup (ethnicity, English language proficiency, disability) differences in subscores and overall scores were conducted. No significant differences were identified in the overwhelming majority of analyses. Table 9 (p.14) of the FAST Response to the Assessment Design Standards outlines a two-year review cycle for future differential effects analyses. Information about psychometric analyses is provided in Appendix G of the FAST Response to the Assessment Design Standards document.
1(l) In designing assessment administration procedures, the model sponsor includes administrative accommodations that preserve assessment validity while addressing issues of access for candidates with disabilities or learning needs.	4	The FASTv2.0 Complete Manual (p.41) states that "candidates with disabilities will be reasonably accommodated" and should "contact their University Coaches and the University Services for Students with Disabilities" who will coordinate with faculty and staff to assist in providing these. It further clarifies that candidates' responses to the tasks must reflect their "own unaided work." The FAST Response to the Assessment Design Standards (p.14) notes that "candidates can seek assistance prior to completing the task to ensure concept understanding and may use assistive devices as appropriate to help them complete the task." At this time, there is no evidence to demonstrate comparability of scores on accommodated assessments. FAST may be too small to allow for quantitative analysis of comparability on accommodated assessments; however, even qualitative information from relatively small groups could help to bolster support for preserving assessment validity.

(continued)

Table 1.2. (Continued)

Assessment Design Standard	FAST Rating	Rationale for FAST Rating
1(m) See Chapter 5		
1(n) To preserve the validity and fairness of the assessment over time, the model sponsor may need to develop and field test new pedagogical assessment tasks and multi-level scoring rubrics to replace or strengthen prior ones. Initially and periodically, the model sponsor analyzes the assessment tasks and scoring rubrics to ensure that they yield important evidence that represents candidate knowledge and skill related to the TPEs, and serve as a basis for determining entry-level pedagogical competence to teach the curriculum and student population of California's TK-12 public schools. The model sponsor documents the basis and results of each analysis, and modifies the tasks and rubrics as needed.	5	During the development and field testing of FAST 2.0 tasks, input was solicited from a variety of educators, including university faculty and program supervisors, master teachers, and local support providers of new teachers to ensure that revised tasks and rubrics adequately measure the entry-level pedagogical competence of teacher candidates as articulated in the revised TPEs. In addition, the educators provided input on the appropriateness of an initial passing standard. Results of data analyses from field tests as well as recommendations from scorers, faculty, and candidates resulted in minor clarifications in wording of directions and rubrics" (FAST Response to the Assessment Design Standards, p.15). Analyses of pass rates, scoring reliability, and subgroup (ethnicity, English language proficiency, disability) differences in subscores and overall scores were conducted. No significant differences were identified in the overwhelming majority of analyses. Table 9 in the FAST response (p.14) outlines a two-year review cycle for future differential effects analyses. Information about psychometric analyses is provided in Appendix G of the FAST Response to the Assessment Design Standards.
1(o) The model sponsor must make all TPA materials available to the Commission upon request for review and approval, including materials that are proprietary to the model sponsor. The Commission will maintain the confidentiality of all materials designated as proprietary by the model sponsor.	5	The FAST Response to the Assessment Design Standards states that the model sponsor for the FAST will "continue to make all TPA materials available to the Commission as requested (p.16)." Moreover, upon request, the FAST model sponsor has made all requested materials available to the researchers conducting this comparability investigation, which can be shared with the Commission.
2(a) See Chapter 4		

(continued)

Table 1.2. (Continued)

Assessment Design Standard	FAST Rating	Rationale for FAST Rating
2(b) Pedagogical assessment tasks and scoring rubrics are extensively field tested in practice before being used operationally in the Teaching Performance Assessment. The model sponsor evaluates the field test results thoroughly and documents the field test design, participation, methods, results and interpretation.	5	Information about field testing is provided in the Update on the Redevelopment of the Fresno Assessment of Student Teachers (FAST) and Request for Commission Authority to Waive the Professional Preparation Requirement for Candidates Participating in the FAST Field Test (4D FAST Waivers 082417.doc), and in Appendix G of the FAST Response to the Assessment Design Standards. The FAST Waivers document describes the date range of field testing and the number and characteristics (single subject or multiple subject) of candidates participating. The FAST Response to the Assessment Design Standards states that the Site Visitation Project was field-tested in Fall 2017 and again during the Spring 2018 semester. The Teaching Sample Project was field-tested during the Spring 2018 semester. Field tests were designed to evaluate the validity and reliability of the assessment. Appendix G presents information about psychometric analyses, including analyses of pass rates, scoring reliability, and subgroup (ethnicity, English language proficiency, disability) differences in scores.
2(c) See Chapter 4		
2(d) In conjunction with the provisions of the applicable Teacher Preparation Program Standards relating to the Teaching Performance Assessment, the model 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.	5	Evaluation of the effectiveness of scorer training procedures is built into the two-year review cycle of tasks and training (FAST Response to the Assessment Design Standards, p.14). According to the response document, “When scorers have completed their task scoring, they will provide the FAST Coordinator with comments and recommendations via a written evaluation of how well the training session they attended prepared them to score candidates’ work using the scoring rubrics. Specific information requested will include the degree to which they understand the TPE elements being evaluated by the task; the degree to which they understand the directions that candidates must follow to successfully complete the task; the degree to which they understand the qualitative descriptors at each level of the task’s scoring rubric; and their own biases that may unfairly affect scoring. In addition to input from scorers, reliability data will be used to determine consistency in scoring. Higher reliability suggests more effective training. Surveys and data analysis results will be reviewed and updates to the training program will be made as indicated” (p.19).
2(e) See Chapter 4		

(continued)

Table 1.2. (Continued)

Assessment Design Standard	FAST Rating	Rationale for FAST Rating
2(f) The model sponsor's assessment design includes a clear and easy to implement 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, if the program is using centralized scoring provided by the model sponsor. If the program is implementing a local scoring option, the program must provide an appeal process as described above for candidates who do not pass the assessment. Model sponsors must document that all candidate appeals granted a second scoring are scored by a new assessor unfamiliar with the candidate or the candidate's response.	4	The Complete Manual (p.42) outlines the appeal policy, and Appendix C of the FAST Response to the Assessment Design Standards provides further detail. Appealed scores are reviewed by a panel of three trained scorers who were not involved in the original scoring. No documentation was provided showing that new assessors, unfamiliar with the candidate or candidate's response, conducted the second scoring.
2(g) See Chapter 4		
2(h) The model sponsor provides program level aggregate results to the Commission, in a manner, format and time frame specified by the Commission, as one means of assessing program quality. It is expected that these results will be used within the Commission's ongoing accreditation system.	5	Appendix G of the FAST Response to the Assessment Design Standards reports overall pass rates, based on field test data. The response document also states that the model sponsor "will continue to provide aggregate results to the commission" (p.21).
3(a) The model sponsor provides technical assistance to programs implementing the model to support fidelity of implementation of the model as designed. Clear implementation procedures and materials such as a candidate and a program handbook are provided by the model sponsor to programs using the model.	5	The FAST Response to the Assessment Design Standards indicates that the "FAST Coordinator monitors the technical assistance provided to the Multiple and Single Subject programs to ensure both programs implement the model as designed" (p.22). The FAST v2.0 Complete Manual and the Additional Resources for Teacher Candidates provide clear descriptions of procedures and example materials. The FAST Coordinator also conducted seminars with candidates and scorers to ensure a common understanding of FAST.
3(b) A model sponsor conducting scoring for programs is responsible for providing TPA outcomes data at the candidate and program level to the program within three weeks and to the Commission, as specified by the Commission. The model sponsor supervising/moderating local program scoring oversees data collection, data review with programs, and reporting.	5	The FAST Response to the Assessment Design Standards states that "All outcomes data from the FAST tasks are provided to the Multiple and Single Subject Programs as well as to the candidates within three weeks of the task submission. The FAST Coordinator, with the assistance of the faculty member responsible for psychometric studies and data analysis, oversees data collection, review with programs, and reporting (p.22)." In 2018–19, the Commission did not require an annual report.

(continued)

Table 1.2. (Continued)

Assessment Design Standard	FAST Rating	Rationale for FAST Rating
3(c) The model sponsor is responsible for submitting at minimum an annual report to the Commission describing, among other data points, the programs served by the model, the number of candidate submissions scored, the date(s) when responses were received for scoring, the date(s) when the results of the scoring were provided to the preparation programs, the number of candidate appeals, first time passing rates, candidate completion passing rates, and other operational details as specified by the Commission.	NA	The Commission is not requiring an annual report at this time.
3(d) The model sponsor is responsible for maintaining the currency of the TPA model, including making appropriate changes to the assessment tasks and/or to the scoring rubrics and associated program, candidate, and scoring materials, as directed by the Commission when necessitated by changes in TK-12 standards and/or in teacher preparation standards.	5	The FAST has undergone a number of changes in response to the revised ADS, as demonstrated by the Transition Plan, the FAST 2.0 changes pdf, and the overview of FAST 2.0 presented at the TAC meeting on November 16, 2017. These serve as evidence that the model sponsor is maintaining the currency of FAST.
3(e) The model sponsor must define the retake policies for candidates who fail one or more parts of the TPA which preserve the reliability and validity of the assessment results. The retake policies must include whether the task(s) on which the candidate was not successful must be retaken in whole or in part, with appropriate guidance for programs and candidates about which task and/or task components must be resubmitted for scoring by a second assessor and what the resubmitted response must include.	5	The FASTv2.0 Complete Manual describes the non-passing score procedure on p.41-42. The FAST Response to the Assessment Design Standards describes actions in case of non-passing score on p.23.

Note. NA = Not applicable.

edTPA. Table 1.3 presents the ratings for edTPA on each ADS. Based on the documentation and materials provided for edTPA, we were able to provide ratings on all but one ADS. All Standards were rated a '4' (Evidence in the documentation mostly covers all aspects of the Standard/element) or '5' (Evidence in the documentation fully covers all aspects of the Standard/element), with the majority receiving a rating of a '5' on the 5-point rating scale. The average rating on the ADS for the edTPA was 4.76, which is a slight increase from the average Year 1 rating of 4.64.

Table 1.3. Ratings on the Assessment Design Standards for edTPA

Assessment Design Standard	edTPA Rating	Rationale for edTPA Rating
1(a) See Chapter 4		
1(b) The TPA model sponsor must include a focus on content-specific pedagogy within the design of the TPA tasks and scoring scales to assess the candidate's ability to effectively teach the content area(s) authorized by the credential.	5	Each edTPA handbook embeds a subject-specific focus into a common architecture addressing the integration of planning, instruction, and assessment. Candidates are required to support student learning of the knowledge and skills within that subject area. In the "edTPA Subject Specific Handbook Foci" (see the Transition Plan), the model sponsor outlined the content-specific pedagogy for each subject-specific area of the edTPA assessments. For scoring scales, as mentioned in the Transition Plan, rubric 9 is a subject-specific rubric designed to assess subject-specific constructs identified by the design team for each content area. Six rubrics (1, 5, 7, 8, 9, and 15; 1, 4, 6, 7, 8, and 14 in WL; and CL) have subject-specific language embedded within the rubric criteria and evidence for rubrics is interpreted by scorers through a subject-specific lens. There are also two versions of the elementary education handbook—one with literacy focus and one with mathematics focus, each with 18 rubrics.
1(c) See Chapter 4		
1(d) The model sponsor must include within the design of the TPA candidate tasks a focus on addressing the teaching of English learners, all underserved education groups or groups that need to be served differently, and students with special needs in the general education classroom to adequately assess the candidate's ability to effectively teach all students.	5	Across edTPA tasks, candidates were directed to consider the variety of learners in their class who might require different strategies/support (e.g., students with IEPs or 504 plans, English language learners, struggling readers, underperforming students or those with gaps in academic knowledge, and/or gifted students). Furthermore, in order to meet TPA Design Standard 1(d), edTPA handbooks now include a revised note for California candidates indicating that candidates must select focus students that meet the following requirements: "English learner, underserved education groups, and a student with specific learning need."
1(e) For Multiple Subject candidates, the model sponsor must include assessments of the core content areas of at least Literacy and Mathematics. Programs use local program performance assessments for History-Social Science and Science if not already included as part of the TPA.	5	As described in the Transition Plan, to meet the request of programs and the TPA Design Standards, edTPA developed an additional version of the Elementary Education Handbook, whereby Tasks 1-3 address candidates' performance in Elementary Mathematics (i.e., the Elementary Mathematics Handbook) and Task 4 assesses candidate performance on Elementary Literacy. Task 4 for Elementary Literacy is similar to the Elementary Mathematics Task 4; the handbook includes an additional three rubrics that focus on the candidates' ability to assess students' literacy and to plan and support the re-engagement of students in a focused learning experience. The new Elementary Education (mathematics focus) Handbook was field tested in five California teacher preparation programs during the 2017–18 program year and was found to be comparable to previous handbook tasks. Programs use local assessments for History-Social Science and Science.

(continued)

Table 1.3. (Continued)

Assessment Design Standard	edTPA Rating	Rationale for edTPA Rating
1(f) The model sponsor must include a focus on classroom teaching performance within the TPA, including a video of the candidate's classroom teaching performance with candidate commentary describing the lesson plan and rationale for teaching decisions shown and evidence of the effect of that teaching on student learning.	5	As described in the Transition Plan, the evidence collected for edTPA portfolios include video clips of instruction, lesson plans, student work samples, analysis of student learning, and reflective commentaries analyzing and justifying instructional decisions based on candidates' understandings of varied student strengths and needs. Video commentaries require candidates to explain a) how lesson plans are designed with students in mind, b) the effectiveness of the instruction, and c) the impact of the lesson on student learning.
1(g) See Chapter 4		
1(h) See Chapter 4		
1(i) The model sponsor provides a clear statement acknowledging the intended uses of the assessment. The statement demonstrates the model sponsor's clear understanding of the implications of the assessment for candidates, preparation programs, the public schools, and TK-12 students. The statement includes appropriate cautions about additional or alternative uses for which the assessment is not valid. All elements of assessment design and development are consistent with the intended uses of the assessment for determining the pedagogical competence of candidates for Preliminary Teaching Credentials in California and as information useful for determining program quality and effectiveness.	4	In the Transition Plan, the model sponsor provided a clear statement of the intended uses of the assessment. The intended uses of the assessment include (1) determining the pedagogical content competency of candidates for Preliminary Teaching Credentials in California, as well as (2) providing information useful for determining program quality and effectiveness. The elements of assessment design and development are consistent with the intended uses of the assessment. In the Transition Plan, the model sponsor described that edTPA outcomes are intended to be used by state agencies and teacher preparation programs for purposes such as serving as an assessment of pre-service teaching; augmenting existing assessments of candidates for teacher licensure that focus on basic skills and/or subject-matter knowledge; producing information to be used for state and national accreditation of teacher preparation programs and/or program completion, and providing actionable evidence to guide decision-making about teacher preparation program revision and improvement. The model sponsor cautioned that edTPA scores should be used in combination with other measures of performance. The completion and passing of edTPA alone was insufficient to demonstrate a candidate's qualifications to become a teacher of record. There is no statement in the edTPA documentation regarding implications of edTPA for public schools and TK-12 students. Score reports are accompanied by an interpretive guide to help candidates and programs make the best use of the detailed information provided. See Interpreting your edTPA score profile here: http://www.edtpa.com/Content/Docs/edTPA_InterpretingYourProfile.pdf Program reporting is described here: http://www.edtpa.com/PageView.aspx?f=GEN_edTPAReporting.html Program file layout here: http://www.edtpa.com/Content/Docs/edTPA_InstitutionReportLayout.pdf

(continued)

Table 1.3. (Continued)

Assessment Design Standard	edTPA Rating	Rationale for edTPA Rating
1(j) The model 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.	5	In the Transition Plan (p.60), the model sponsor provided an overview of the bias review process. In the FT summary report (p.441 of the Transition Plan), the model sponsor also provided an overview of the Bias and Sensitivity Review. The Bias Review Meeting Orientation Manual outlines the review procedures and guidelines. The Global Handbook Bias Report presents comments and suggestions from the Bias Review Committee's review of the Assessment Handbook. Comments and suggestions generally called for review for clarity, consistency, complexity, or completion, and were tagged by bias type (e.g., language, fairness, inclusion, content, stereotype).
1(k) The model sponsor completes initial and periodic basic psychometric analyses to identify pedagogical assessment tasks and/or scoring rubrics that show differential effects in relation to candidates' race, ethnicity, language, gender or disability. When group pass-rate differences are found, the model sponsor investigates the potential sources of differential performance and seeks to eliminate construct-irrelevant sources of variance.	4	As described in the Transition Plan, the model sponsor completed initial and periodic basic psychometric analyses to identify subgroup differences. Analyses were conducted for subgroups based on the 2012 and 2013 field test and 2015 operational administration. The model sponsor cautioned about the generalizability of the subgroup results because of several factors such as the small sample sizes of some of the subgroups. The model sponsor mentioned conducting additional research to better understand the differences in subgroup performance but hadn't conducted research in this area yet.
1(l) In designing assessment administration procedures, the model sponsor includes administrative accommodations that preserve assessment validity while addressing issues of access for candidates with disabilities or learning needs.	4	As described in the Transition Plan and the edTPA website, administration procedures are established for edTPA that accommodate candidates requiring alternative arrangements. The model sponsor described that "Administration procedures are established for edTPA that accommodate candidates requiring alternative arrangements while at the same time preserving assessment validity." A Requesting Alternative Arrangements webpage notes that candidates may use screen reading software and/or a scribe to complete the required documents, without advance notice, candidates are also required to attest that they are the sole author of their submission. Other alternative arrangements must be requested, with supporting documentation submitted along with the request. Supporting documentation must provide evidence of review by a qualified professional, must recommended alternative arrangements that are clearly related to the identified disabilities or learning needs, and must be current. There was no documentation of evidence to demonstrate comparability of scores on accommodated assessments. Even qualitative information from relatively small groups could help to bolster support for preserving assessment validity on accommodated assessments.

(continued)

Table 1.3. (Continued)

Assessment Design Standard	edTPA Rating	Rationale for edTPA Rating
1(m) See Chapter 5		
1(n) To preserve the validity and fairness of the assessment over time, the model sponsor may need to develop and field test new pedagogical assessment tasks and multi-level scoring rubrics to replace or strengthen prior ones. Initially and periodically, the model sponsor analyzes the assessment tasks and scoring rubrics to ensure that they yield important evidence that represents candidate knowledge and skill related to the TPEs, and serve as a basis for determining entry-level pedagogical competence to teach the curriculum and student population of California's TK-12 public schools. The model sponsor documents the basis and results of each analysis, and modifies the tasks and rubrics as needed.	5	The assessment was field tested in 2012 and 2013 (See the 2013 edTPA Field Test: Summary Report), and additional elements to align with TPEs were implemented in 2017–18 (see CA edTPA Cover Letter, May 2018). Field test results were used to improve the assessment. Appendix 1 of the edTPA Transition Plan provided a crosswalk detailing the alignment between edTPA tasks and rubrics to the TPEs to demonstrate that the tasks and the scoring rubrics yielded evidence that represented candidate knowledge and skill related to the TPEs. Additionally, as described in the Transition Plan, to meet the request of programs and the Assessment Design Standards, edTPA developed an additional version of the Elementary Education Handbook, whereby Tasks 1-3 address candidates' performance in Elementary Mathematics (i.e., the Elementary Mathematics Handbook) and Task 4 assesses candidate performance on Elementary Literacy.
1(o) The model sponsor must make all TPA materials available to the Commission upon request for review and approval, including materials that are proprietary to the model sponsor. The Commission will maintain the confidentiality of all materials designated as proprietary by the model sponsor.	5	In the Transition Plan, the model sponsor explained that any materials designated as proprietary by the model sponsor would be clearly identified for the Commission in order to maintain those materials as confidential. edTPA materials were hosted by either AACTE or Evaluation Systems online sites, and access could be provided to the Commission upon request. Moreover, upon request, the model sponsor has made all requested materials available to the researchers conducting this comparability investigation, which can be shared with the Commission.
2(a) See Chapter 4		
2(b) Pedagogical assessment tasks and scoring rubrics are extensively field tested in practice before being used operationally in the Teaching Performance Assessment. The model sponsor evaluates the field test results thoroughly and documents the field test design, participation, methods, results and interpretation.	5	Field tests were conducted in 2012 and 2013 to improve key aspects of edTPA assessment instruments and supports. As shown in the 2013 edTPA Field Test: Summary Report, the field test results were evaluated, and the model sponsor documented the field test design, participation, methods, results and interpretation in detail.
2(c) See Chapter 4		
2(d) In conjunction with the provisions of the applicable Teacher Preparation Program Standards relating to the Teaching Performance Assessment, the model 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.	5	In the Transition Plan, the model sponsor indicated that an annual review of the scorer training was implemented, which included quantitative and qualitative feedback from scorers, scoring trainers, and scoring supervisors. As a result of this feedback, scorer training modules had been improved and delivered to scorers. The edTPA Training Improvement Timeline details updates that have been made to the training process from Fall 2015 to the present.

(continued)

Table 1.3. (Continued)

Assessment Design Standard	edTPA Rating	Rationale for edTPA Rating
2(e) See Chapter 4		
2(f) The model sponsor's assessment design includes a clear and easy to implement 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, if the program is using centralized scoring provided by the model sponsor. If the program is implementing a local scoring option, the program must provide an appeal process as described above for candidates who do not pass the assessment. Model sponsors must document that all candidate appeals granted a second scoring are scored by a new assessor unfamiliar with the candidate or the candidate's response.	4	The appeal procedure is clearly described on the edTPA website. As indicated in the Transition Plan, the appeals process involves having a scoring supervisor or trainer, one who did not serve as one of the original scorers, review the portfolio submission and original reported scores to confirm the accuracy of the scores provided. No documentation was provided showing that new assessors, unfamiliar with the candidate or candidate's response, conducted the second scoring.
2(g) See Chapter 4		
2(h) The model sponsor provides program level aggregate results to the Commission, in a manner, format and time frame specified by the Commission, as one means of assessing program quality. It is expected that these results will be used within the Commission's ongoing accreditation system.	5	In the Transition Plan, the model sponsor describes the edTPA ResultsAnalyzer® system and the semi-annual summary reports, and that results are provided to programs and the Commission. In the Transition Plan, the model sponsor explains that the ResultsAnalyzer® is aligned to the structure, components, and reporting of edTPA to support data of performance results. The edTPA State Performance Summary and edTPA National Performance Summary present mean candidate performance, an abbreviated total score distribution, and distributions of rubric level scores by field, to assist programs and state agencies in evaluating their candidates' performance relative to others within the state or across the nation. Recipients of these summary reports also receive a Read Me file that contains suggested uses for each report section.
3(a) The model sponsor provides technical assistance to programs implementing the model to support fidelity of implementation of the model as designed. Clear implementation procedures and materials such as a candidate and a program handbook are provided by the model sponsor to programs using the model.	5	As described in the Transition Plan, the model sponsor provides various supports to programs implementing the model (e.g., online community, resource library, edTPA website, Online Platform for Preparation Programs). edTPA's National Academy of consultants also provides on-site professional development and implementation support. The model sponsor provides participating teacher preparation programs with a wide array of support materials and professional development for faculty and P–12 partners. Clear implementation procedures and materials are provided to both candidates and faculties in documents such as the edTPA Making Good Choices, the edTPA Handbooks, and the guidelines for supporting candidates.

(continued)

Table 1.3. (Continued)

Assessment Design Standard	edTPA Rating	Rationale for edTPA Rating
3(b) A model sponsor conducting scoring for programs is responsible for providing TPA outcomes data at the candidate and program level to the program within three weeks and to the Commission, as specified by the Commission. The model sponsor supervising/moderating local program scoring oversees data collection, data review with programs, and reporting.	5	The edTPA website indicates that candidates and any program or state agency that the candidate indicates will receive scores three weeks after assessments are submitted. The model sponsor has been utilizing both centralized and local (regional) scoring methodologies, and for both modes of scoring has overseen all scoring data collection, review, and reporting to candidates, programs, and the Commission. Regional scorers complete the same training and qualify using the same criteria before scoring and have the same quality monitoring and scoring consistency requirements as used for centralized scoring.
3(c) The model sponsor is responsible for submitting at minimum an annual report to the Commission describing, among other data points, the programs served by the model, the number of candidate submissions scored, the date(s) when responses were received for scoring, the date(s) when the results of the scoring were provided to the preparation programs, the number of candidate appeals, first time passing rates, candidate completion passing rates, and other operational details as specified by the Commission.	NA	The Commission is not requiring an annual report at this time.
3(d) The model sponsor is responsible for maintaining the currency of the TPA model, including making appropriate changes to the assessment tasks and/or to the scoring rubrics and associated program, candidate, and scoring materials, as directed by the Commission when necessitated by changes in TK-12 standards and/or in teacher preparation standards.	5	The comprehensive Transition Plan provides evidence that the model sponsor made changes as directed by the Commission to maintain the currency of the TPA model. Data from edTPA are reviewed annually to inform whether changes are needed to handbook directions, assessment tasks, rubrics, or score scales. Changes are also made in response to changes in state policy.
3(e) The model sponsor must define the retake policies for candidates who fail one or more parts of the TPA which preserve the reliability and validity of the assessment results. The retake policies must include whether the task(s) on which the candidate was not successful must be retaken in whole or in part, with appropriate guidance for programs and candidates about which task and/or task components must be resubmitted for scoring by a second assessor and what the resubmitted response must include.	5	The retake policies on the edTPA website provide instructions for candidates to retake either all or part of the assessment. The Retake Instructions for Candidates (https://www.edtpa.com/Content/Docs/edTPATaskRetakeInstructions.pdf) include specifications detailing the artifacts and commentaries for planning, instruction, or assessment of student learning that may or may not be resubmitted as part of a full- or partial-assessment retake.

Note. NA = Not applicable.

CalTPA. Table 1.4 presents the ratings for CalTPA on each ADS. Similar to FAST, considerably more technical documentation was available for CalTPA in Year 2 than in Year 1. In Year 1, the average rating across the ADS for which there was available technical documentation was 4.57. In Year 2, all but one ADS was rated, and the average rating increased to 4.83. All Standards were rated a '4' (Evidence in the documentation mostly covers all aspects of the Standard/element) or '5' (Evidence in the documentation fully covers all aspects of the Standard/element), with the majority receiving a rating of a '5' on the 5-point rating scale.

Table 1.4. Ratings on the Assessment Design Standards for CalTPA

Assessment Design Standard	CalTPA Rating	Rationale for CalTPA Rating
1(a) See Chapter 4		
1(b) The TPA model sponsor must include a focus on content-specific pedagogy within the design of the TPA tasks and scoring scales to assess the candidate's ability to effectively teach the content area(s) authorized by the credential.	5	The CalTPA Assessment Guides state explicitly that the model is structured to address "developmentally appropriate practices in relation to content specific pedagogy." Instructional Cycle 1 for requires candidates to develop and teach one content-specific lesson within a school placement. In Instructional Cycle 2, candidates are required to teach a lesson "using the content-specific pedagogy." Associated rubrics for both Instructional Cycles reflect an emphasis on content-specific instructional strategies, learning activities, and assessments. There are content-specific rubrics for Instructional Cycle 2 (Assessment-Driven Instruction). The Update on the Redevelopment of the CalTPA describes convening content experts from each subject area to review the Instructional Cycles and rubrics. CalTPA Assessors are required to have content area expertise.
1(c) See Chapter 4		
1(d) The model sponsor must include within the design of the TPA candidate tasks a focus on addressing the teaching of English learners, all underserved education groups or groups that need to be served differently, and students with special needs in the general education classroom to adequately assess the candidate's ability to effectively teach all students.	5	Candidates are required to identify three focus students that address these groups and are evaluated on their ability to address their specific learning needs. Specific rubrics evaluate instruction as it pertains to each focus student.
1(e) For Multiple Subject candidates, the model sponsor must include assessments of the core content areas of at least Literacy and Mathematics. Programs use local program performance assessments for History-Social Science and Science if not already included as part of the TPA.	5	Multiple Subject candidates must demonstrate both literacy and mathematics instruction within the CalTPA. Candidates are offered two alternate approaches. In the first, they can focus on either literacy or mathematics in Cycle 1, and then on the other area in Cycle 2. In the second, they can focus one of the two cycles on integrating literacy with another content area(s), and then focus the other cycle on integrating mathematics with another content area(s).
1(f) The model sponsor must include a focus on classroom teaching performance within the TPA, including a video of the candidate's classroom teaching performance with candidate commentary describing the lesson plan and rationale for teaching decisions shown and evidence of the effect of that teaching on student learning.	5	Candidates are required to video-record instruction in both Instructional Cycles. Candidates provide commentary by selecting and annotating clips from the videos. Annotations are to provide descriptions and rationales for how and why they approached teaching to specific learning goals and objectives, provided content-specific feedback to students, monitored student content learning and development of academic learning, selected the assessment strategies used, and chose the strategies used to establish a positive and safe learning environment. Candidates are then to reflect on assessment results.
1(g) See Chapter 4		

(continued)

Table 1.4. (Continued)

Assessment Design Standard	CalTPA Rating	Rationale for CalTPA Rating
1(h) See Chapter 4		
1(i) The model sponsor provides a clear statement acknowledging the intended uses of the assessment. The statement demonstrates the model sponsor's clear understanding of the implications of the assessment for candidates, preparation programs, the public schools, and TK-12 students. The statement includes appropriate cautions about additional or alternative uses for which the assessment is not valid. All elements of assessment design and development are consistent with the intended uses of the assessment for determining the pedagogical competence of candidates for Preliminary Teaching Credentials in California and as information useful for determining program quality and effectiveness.	4	The CalTPA Performance Assessment Overview (Version 02) document states that the CalTPA is one of multiple measures to inform candidate preparedness. It goes on to state that the CalTPA is intended to provide both a formal assessment of candidate ability and a framework of performance-based guidance to inform candidate preparation and continued professional growth. Furthermore, it states that feedback provided at the completion of each cycle is intended to facilitate preparation for the subsequent assessment cycle and that data is shared with institutions to assist them in making program improvements and to guide induction programs as they work with new teachers to individualize learning plans. The CalTPA is intended to provide authentic evidence of teaching ability and student learning experienced during clinical practice. There is no statement in the CalTPA documentation regarding the implications of CalTPA for public schools and TK-12 students. The Candidate Score Report states that the Results Report "is for your records only" and that, "This assessment was not designed to compare your performance to that of other candidates. Your score is used to compare your performance to the performance level set by the Commission on Teacher Credentialing."
1(j) The model 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.	5	Content expert panels were convened in October 2016 to review the CalTPA Instructional Cycles, rubrics, and materials. The CalTPA Bias Review Committee convened in August 2017 to review the draft CalTPA and comment on potential bias issues. Content experts provided comments on subject-specific pedagogy and recommended revisions for the pilot assessment. The Bias Action Summary documents the comments made by the Committee along with actions taken in response and any further follow-up actions. Online Bias Review Conferences were conducted in August 2018. Assessment materials were reviewed by applying several criteria related to content, language, offense, stereotypes, fairness, and diversity. These criteria addressed potential bias due to gender, race, nationality, national origin, ethnicity, religion, age, sexual orientation, disability, and cultural, economic, or geographic background.
1(k) The model sponsor completes initial and periodic basic psychometric analyses to identify pedagogical assessment tasks and/or scoring rubrics that show differential effects in relation to candidates' race, ethnicity, language, gender or disability. When group pass-rate differences are found, the model sponsor investigates the potential sources of differential performance and seeks to eliminate construct-irrelevant sources of variance.	5	The Update on the Redevelopment of the CalTPA document mentions that performance data from the pilot was one of the data sources used by the Design Team in revisions to the scoring rubrics. Participation rates were provided by content area, race/ethnicity, gender, program type, program length, field placement type, and field placement setting. Only eight candidates participating in the pilot did not meet the passing threshold. Additional bias reviews were conducted by the model sponsor in August 2018.

(continued)

Table 1.4. (Continued)

Assessment Design Standard	CalTPA Rating	Rationale for CalTPA Rating
1(l) In designing assessment administration procedures, the model sponsor includes administrative accommodations that preserve assessment validity while addressing issues of access for candidates with disabilities or learning needs.	4	According to the model website, candidates may request for alternate assessment arrangements due to a diagnosed disability. According to the policy, "it is acceptable for a candidate to use screen reading software and/or a scribe to complete submissions." All candidates are required to sign an attestation that they are the sole authors of their task responses. Candidates are instructed to submit a letter providing the specifics of their requested alternate arrangements. Required documentation that must accompany the request includes a current, signed statement from the diagnosing professional, an indication of the diagnosis, and recommended alternative arrangements, as well as other documentation to show either a history of special education services, psychological test results, or medical test results. At this time, there is no evidence to demonstrate comparability of scores on accommodated assessments. Even qualitative information from relatively small groups could help to bolster support for preserving assessment validity on accommodated assessments.
1(m) See Chapter 5		
1(n) To preserve the validity and fairness of the assessment over time, the model sponsor may need to develop and field test new pedagogical assessment tasks and multi-level scoring rubrics to replace or strengthen prior ones. Initially and periodically, the model sponsor analyzes the assessment tasks and scoring rubrics to ensure that they yield important evidence that represents candidate knowledge and skill related to the TPEs, and serve as a basis for determining entry-level pedagogical competence to teach the curriculum and student population of California's TK-12 public schools. The model sponsor documents the basis and results of each analysis, and modifies the tasks and rubrics as needed.	5	According to the Update on the Redevelopment of the CalTPA, the CalTPA was piloted in early 2017 and the qualitative and quantitative findings were reviewed by the Design Team, Evaluation Systems Group of Pearson, and the Commission. "The DT [Design Team] provided thoughtful recommendations based on findings" and worked to "revise the cycles, rubrics, and assessment materials in preparation for the field test." Surveys were completed by candidates, program coordinators and assessors who participated in the pilot, and their feedback informed the revisions process reflected in the CalTPA field test. Similar information was gathered from the CalTPA field test in 2017–18 and reflected in the Year 1 operational CalTPA in 2018–19.
1(o) The model sponsor must make all TPA materials available to the Commission upon request for review and approval, including materials that are proprietary to the model sponsor. The Commission will maintain the confidentiality of all materials designated as proprietary by the model sponsor.	5	The Commission is the model sponsor thereby facilitating access to all CalTPA materials. Moreover, upon request, the model sponsor has made all requested materials available to the researchers conducting this comparability investigation, which can be shared with the Commission.

(continued)

Table 1.4. (Continued)

Assessment Design Standard	CalTPA Rating	Rationale for CalTPA Rating
2(a) See Chapter 4		
2(b) Pedagogical assessment tasks and scoring rubrics are extensively field tested in practice before being used operationally in the Teaching Performance Assessment. The model sponsor evaluates the field test results thoroughly and documents the field test design, participation, methods, results and interpretation.	5	Pilot testing was done to inform field testing; its design, participation, methods, and results are well-documented. The PowerPoint presentation from the CalTPA Design Team meeting (July 2018) describes the field test sample, including the number of candidates, number of programs, number of submissions, submission rate, and the gender, racial, geographic, and placement characteristics. Average performance on each rubric, correlations between rubrics, and factor analysis results were also reported. Finally, post-test survey and focus group results were presented. Additional information is included in the "Update on the Redevelopment of the CalTPA" document. Plans for review and revisions in response to field test results are discussed.
2(c) See Chapter 4		
2(d) In conjunction with the provisions of the applicable Teacher Preparation Program Standards relating to the Teaching Performance Assessment, the model 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.	5	The CalTPA collected feedback from Assessors and Lead Assessors via surveys (both Field Test and Operational Year 1). They were asked to provide feedback about the scoring process, including ratings of the clarity of performance levels, the sufficiency of evidence for scoring, and their confidence in the scores they assigned. The "Update on the Redevelopment of the CalTPA" document states that "Lead assessors will continue to work with Commission staff and ES over the summer to revise assessor training and to prepare materials for the fall, online and in-person assessor trainings" (p.10).
2(e) See Chapter 4		
2(f) The model sponsor's assessment design includes a clear and easy to implement 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, if the program is using centralized scoring provided by the model sponsor. If the program is implementing a local scoring option, the program must provide an appeal process as described above for candidates who do not pass the assessment. Model sponsors must document that all candidate appeals granted a second scoring are scored by a new assessor unfamiliar with the candidate or the candidate's response.	4	As stated in the CalTPA Scoring Quality Management Plan, candidates may request a score verification if they do not pass. Lead assessors review these submissions to verify if the initially reported scores are accurate. If the lead identifies one or more scores that are higher than initially reported, revised scores will be reported to the candidate. The process for requesting a score verification is clearly described here: http://www.ctcexams.nesinc.com/PageView.aspx?f=GEN_RequestingARescore.html. No documentation was provided showing that new assessors, unfamiliar with the candidate or candidate's response, conducted the scoring.

(continued)

Table 1.4. (Continued)

Assessment Design Standard	CalTPA Rating	Rationale for CalTPA Rating
2(g) See Chapter 4		
2(h) The model sponsor provides program level aggregate results to the Commission, in a manner, format and time frame specified by the Commission, as one means of assessing program quality. It is expected that these results will be used within the Commission's ongoing accreditation system.	5	Data include overall, cycle-level, and rubric-level means by content area for each program as well as statewide. The Update on the Redevelopment of the California Teaching Performance Assessment states that the Commission receives diagnostic feedback reports within three weeks of a submission date. The Update on the Redevelopment document explains that data will be used to inform program accreditation visits and provide them with insights on how to design programs to support candidate growth and development (p.14).
3(a) The model sponsor provides technical assistance to programs implementing the model to support fidelity of implementation of the model as designed. Clear implementation procedures and materials such as a candidate and a program handbook are provided by the model sponsor to programs using the model.	5	The CalTPA representatives provided numerous program implementation presentations to stakeholders. Moreover, the CalTPA website outlines policies and guidelines for supporting candidates while they are completing the assessment. This webpage also contains links to updates on the assessment, and details about the assessment tasks. The CalTPA website also provides access to assessment materials including Performance Assessment Guides that present the directions for each Cycle, the associated scoring rubrics, and a glossary of terms used. The Assessment Policies section of the CalTPA website provides guidelines for submitting assessment materials for scoring. A Candidate Support Center is available to assist candidates via phone, email, or live chat.
3(b) A model sponsor conducting scoring for programs is responsible for providing TPA outcomes data at the candidate and program level to the program within three weeks and to the Commission, as specified by the Commission. The model sponsor supervising/moderating local program scoring oversees data collection, data review with programs, and reporting.	5	The CalTPA representatives provided numerous program implementation presentations to stakeholders. Moreover, the CalTPA website outlines policies and guidelines for supporting candidates while they are completing the assessment. This webpage also contains links to updates on the assessment, and details about the assessment tasks. The CalTPA website also provides access to assessment materials including Performance Assessment Guides that present the directions for each Cycle, the associated scoring rubrics, and a glossary of terms used. The Assessment Policies section of the CalTPA website provides guidelines for submitting assessment materials for scoring. A Candidate Support Center is available to assist candidates via phone, email, or live chat.
3(c) The model sponsor is responsible for submitting at minimum an annual report to the Commission describing, among other data points, the programs served by the model, the number of candidate submissions scored, the date(s) when responses were received for scoring, the date(s) when the results of the scoring were provided to the preparation programs, the number of candidate appeals, first time passing rates, candidate completion passing rates, and other operational details as specified by the Commission.	NA	The Commission is not requiring an annual report at this time.

(continued)

Table 1.4. (Continued)

Assessment Design Standard	CalTPA Rating	Rationale for CalTPA Rating
3(d) The model sponsor is responsible for maintaining the currency of the TPA model, including making appropriate changes to the assessment tasks and/or to the scoring rubrics and associated program, candidate, and scoring materials, as directed by the Commission when necessitated by changes in TK-12 standards and/or in teacher preparation standards.	5	The CalTPA has been redeveloped in response to adoption of revised TPAs and TPEs. Changes have been made to tasks and rubrics, as well as program and candidate guidance and scoring models. The revised model was piloted in 2016–17 and field tested in 2017–18. Based on qualitative and quantitative data gathered from pilot and field test, enhancements were made.
3(e) The model sponsor must define the retake policies for candidates who fail one or more parts of the TPA which preserve the reliability and validity of the assessment results. The retake policies must include whether the task(s) on which the candidate was not successful must be retaken in whole or in part, with appropriate guidance for programs and candidates about which task and/or task components must be resubmitted for scoring by a second assessor and what the resubmitted response must include.	5	The Update on the Redevelopment of the California Teaching Performance Assessment states that candidates who do not achieve a performance level of 2 across all rubrics and who have more than one rubric-level score of 1 may receive coaching and support to retake all or part of the failed cycle(s) (p.14). The February 2019 Virtual Think Tank states that the number of retakes must be at least one but may be more depending on local program policies. Resubmissions will be rescored by a different assessor.

Evaluation of Strength of Evidence for Test Design and Development Standards from the Joint Standards

Next, we present the results of the numeric ratings assigned to each of the test design and development Standards from the *Joint Standards.* The following *Joint Standards* are not applicable to the TPA models and thus are not included: 4.3, 4.4, 4.11, 4.14, 4.17, 4.19, 4.21 and 4.23. The results are presented in Tables 1.5 – 1.7 for FAST, edTPA, and CalTPA, respectively. Each table includes the (a) test design and development Standards from the *Joint Standards* in the left column, (b) rating on the strength of evidence for the Standard in the middle column, and (c) rationale for the rating in the right column.

FAST. Table 1.5 presents the ratings for FAST on each test design and development Standard from the *Joint Standards*. In Year 1 (2017–18), the technical documentation for FAST was sparse, and, thus, many of the Standards could not be rated due to unavailable technical documentation. Since that time, the technical documentation has improved. Thus, all of the applicable test design and development *Joint Standards* were rated in Year 2 (2018–19). The average rating across the *Joint Standards* also increased notably from 3.80 in Year 1 to 4.64 in Year 2. All of the Standards were rated a '4' (Evidence in the documentation mostly covers all aspects of the Standard/element) or '5' (Evidence in the documentation fully covers all aspects of the Standard/element), with the majority receiving a rating of a '5.'

Table 1.5. Ratings on the Joint Standards for FAST

Test Design Standards from the *Joint Standards*	FAST Rating	FAST Rationale
4.1 Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	5	The FAST Tasks Matrix documents the domain measured. The FASTv2.0 Complete Manual describes the intended examinee population (Fresno State Teacher Candidates) and intended uses (to evaluate mastery of TPEs and to meet requirements for receiving credit for the fieldwork course). Information on the alignment between the TPEs and assessment tasks provides a rationale for the intended use. There is no "test specifications" or "test blueprint" document; however, the elements of 4.1 can be found in the aforementioned documentation.
4.2 In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.	4	The FASTv2.0 Complete Manual notes that the FAST is intended to evaluate mastery of the TPEs and provides detail about the test content, including the tasks, directions, and scoring criteria. Procedures for gaining access to accommodations is also provided, and it is emphasized that test responses must reflect candidates' independent work. Appendix G of the FAST Response to the Assessment Design Standards provides psychometric analyses of field test data, including score reliability and pass rates; however, basic descriptive analyses of score data (e.g., rubric level means and standard deviations) are not provided. The model sponsor should consider developing a formal test specifications document to provide the elements of 4.2 in a centralized location, which should include the desired psychometric properties of the test items and test.
4.5 If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	5	The FASTv2.0 Complete Manual (p.41) states that "candidates with disabilities will be reasonably accommodated" and should "contact their University Coaches and the University Services for Students with Disabilities" who will coordinate with faculty and staff to assist in providing these. It further clarifies that candidate's responses to the tasks must reflect their "own unaided work." The FAST Response to the Assessment Design Standards (p.14) notes that "candidates can seek assistance prior to completing the task to ensure concept understanding and may use assistive devices as appropriate to help them complete the task."

(continued)

Table 1.5. (Continued)

Test Design Standards from the *Joint Standards*	FAST Rating	FAST Rationale
4.6 When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.	4	The independent, external evaluators conducting this comparability investigation, all educational researchers with PhDs and five or more years of experience conducting validity studies of educational assessments, meet the intent of this Standard. The reviewers for Activity 1, however, do not represent a demographically diverse group (two females and one male Caucasian).
4.7 The procedures used to develop, review, and try out items and to select items from the item pool should be documented.	5	The Update on the Redevelopment of the FAST, FAST 2.0 Changes, and the FAST Transition Plan provide evidence of the procedures used to develop and review items. The FAST Response to the Assessment Design Standards (p.13) describes a multi-step review process that tasks and rubrics underwent prior to field testing. Appendix G of the FAST Response to the Assessment Design Standards document summarizes analyses of item performance (i.e., pass rates) that identified no concerning subgroup differences. Table 9 of this document (p. 14) also outlines a periodic review schedule that includes analyses of group differences in performance and revisions of directions and rubrics.
4.8 The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.	4	The FAST Response to the Assessment Design Standards (p.13) states that “faculty with expertise in linguistic and culturally sensitivity” consisting of “one Southeast Asian, one African American, one Hispanic, and one Caucasian” reviewed all task components. The FAST Transition Team, composed of “experts in cultural and language foundations, students with special needs, educational psychology, and educational technology” also reviewed the tasks and rubrics. Finally, “two classes of graduate students from diverse backgrounds who had recently completed the teaching credential program at Fresno State reviewed the tasks and rubrics.” Information on the instructions and training given to these reviewers is not available.
4.9 When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.	5	The Update on the Redevelopment of the FAST document describes the planned sample for field testing. Appendix G of the FAST Response to the Assessment Design Standards describes the field test sample including gender, ethnicity, English language fluency, and self-reported disability. The sample was representative of the target population because the field test was administered to the full target population.

(continued)

Table 1.5. (Continued)

Test Design Standards from the *Joint Standards*	FAST Rating	FAST Rationale
4.10 When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.	4	Appendix G of the FAST Response to the Assessment Design Standards describes the field test sample that was used for initial psychometric analyses of the FAST tasks. The field test sample was of adequate size and diversity because the field test sample was the target sample. These analyses included (a) comparisons of score differences among student ethnic, gender, and English language fluency subgroups and (b) reliability among scorers. Model based methods are not used to estimate item parameters for the FAST. Overall scores are pass/no pass decisions based on achieving a minimum score on each element of the two major tasks (SVP and TSP). We recommend that the model sponsor report basic descriptive statistics on rubric scores and overall scores (i.e., means and standard deviations). This information should be used to monitor aspects of tasks that aren't performing as expected (e.g., low scores on particular rubrics, very large standard deviations).
4.12 Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.	5	This is documented in the FAST Tasks Matrix, which maps TPEs to rubrics.
4.13 When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.	5	Potential sources of irrelevant variance include measurement error that is introduced during the scoring process or systematic differences in the observed performances of subgroups that are not related to true differences in their mastery of the TPEs. These potential sources of irrelevant variance were addressed during the development phase though item reviews, and during the scoring phase through scorer training, qualification, and calibration. Appendix G of the FAST Response to the Assessment Design Standards presents evidence that scorers reached high levels of agreement in terms of assigned scores, and that subgroups did not demonstrate substantial performance differences.

(continued)

Table 1.5. (Continued)

Test Design Standards from the *Joint Standards*	FAST Rating	FAST Rationale
4.15 The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.	5	Appendix A of the FAST Response to the Assessment Design Standards presents directions and rubrics for the Site Visitation Project task, and Appendix B presents directions and rubrics for the Teaching Sample Project task. The FASTv2.0 Complete Manual discusses allowable variations in terms of testing accommodations. The provision of accommodations is overseen by the office for University Services for Students with Disabilities, and the FAST Intended Use Policy further clarifies that candidates' responses must reflect unaided student work. The FAST Response to the Assessment Design Standards (p.14) notes that "candidates can seek assistance prior to completing the task to ensure concept understanding and may use assistive devices as appropriate to help them complete the task."
4.16 The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.	5	Appendix A of the FAST Response to the Assessment Design Standards presents directions and rubrics for the Site Visitation Project task, and Appendix B presents directions and rubrics for the Teaching Sample Project task. This document also clearly outlines the scoring rules used to make pass/fail decision. Additional Resources for Teacher Candidates (Available on TK20) provide several templates and examples for both FAST tasks.
4.18 See Chapter 4		
4.20 See Chapter 4		
4.22 See Chapter 4		
4.24 Test specifications should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may reduce the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.	5	The FAST Transition Plan and the FAST Response to the Assessment Design Standards document the revisions to FAST in light of the revised ADS and TPEs.

(continued)

Table 1.5. (Continued)

Test Design Standards from the *Joint Standards*	FAST Rating	FAST Rationale
4.25 When tests are revised, users should be informed of the changes to the specifications, of any adjustments made to the score scale, and of the degree of comparability of scores from the original and revised tests. Tests should be labeled as "revised" only when the test specifications have been updated in significant ways.	4	The Update on the Redevelopment of the Fresno Assessment of Student Teachers (FAST) and Request for Commission Authority to Waive the Professional Preparation Requirement for Candidates Participating in the FAST Field Test document highlights differences between the original FAST and FAST 2.0. The FAST Response document, p.1, describes the FAST as "the *revised* Fresno Assessment of Student Teachers' which is warranted given the scope of the changes described. Two of the four original tasks were carried over into the new version but were modified and/or expanded. These modifications are described in detail in the update document (pp.1-4). The model sponsor intended to maintain the minimum passing standard from the original version (minimum score of 2 on each rubric), though it is not clear if this minimum passing standard was changed as a result of the standard setting reviews that were done. A statement directly addressing the comparability of scores from the two versions would be helpful, particularly for ongoing monitoring of program effectiveness.

edTPA. Table 1.6 presents the ratings for edTPA on each test design and development Standard from the *Joint Standards*. The average rating for edTPA in Year 1 (2017–18) was 4.41, and in Year 2 (2018–19) it increased to 4.77 in light of additional documentation and/or clarifications provide by the model sponsor. All of the Standards were rated a ‘4’ (Evidence in the documentation mostly covers all aspects of the Standard/element) or ‘5’ (Evidence in the documentation fully covers all aspects of the Standard/element), with the majority receiving a rating of a ‘5.’

Table 1.6. Ratings on the Joint Standards for edTPA

Test Design Standards from the *Joint Standards*	edTPA Rating	edTPA Rationale
4.1 Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	5	Per the edTPA website, the examinee population is aspiring teachers and the purpose of edTPA is to measure and support the skills and knowledge that teachers need from Day 1 in the classroom. The edTPA Transition Plan (p.57) provides a statement of the intended uses of the assessment for the intended examinee population, which are: (a) determining the pedagogical content competency of candidates for Preliminary Teaching Credentials in California and (b) providing information useful for determining program quality and effectiveness. The definition of the construct or domain measured can be found in the California Teaching Performance Expectations and edTPA Crosswalk table. The "edTPA technical specifications" document provides an overview of the job analysis survey that was conducted to support the validity of the knowledge, skills, and abilities measured by the edTPA.
4.2 In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.	5	The edTPA technical specifications document describes the job analysis process whereby tasks were rated on their “criticality” and were determined to represent educators’ critical tasks and behaviors. The table on pgs.115-150 of the edTPA Transition Plan depicts content and item formats of the test. The response lengths for submission components are outlined in the Evidence Charts within each handbook. Items are scored via detailed rubrics that differentiate the quality of response at each possible score point. The edTPA Transition Plan describes how factor analysis was used to provide support for the use of a total score (see pg. 22 and Appendix C); the factor analytic results in Appendix C show that the edTPA rubrics loaded on three factors that correspond to the three edTPA tasks (i.e., Planning, Instruction, and Assessment). Descriptive summaries (means and standard deviations) by Task and Rubric are also provided (pg. 295). The edTPA Transition Plan provides a description of scorer reliability and the internal consistency reliability (pg. 271). edTPA also produces State and National Summary reports, which provide mean scores and the distribution of scores at the rubric level and for the total score. The edTPA handbooks provide directions for the test takers and procedures to be used for test administration; additional guidance is available on the edTPA website. Scoring and reporting procedures are described in the edTPA Transition Plan and additional guidance is available on the edTPA website.

(continued)

Table 1.6. (Continued)

Test Design Standards from the *Joint Standards*	edTPA Rating	edTPA Rationale
4.5 If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	5	As described in the Transition Plan and the edTPA website, administration procedures include accommodations for candidates requiring alternative arrangements. The accommodations provided are described on the edTPA website: http://www.edtpa.com/PageView.aspx?f=GEN_RequestingAlternativeArrangements.html. The requirements for permitting the different conditions are documented here: http://www.edtpa.com/PageView.aspx?f=GEN_RequiredDocumentation.html#DocCurrencyPolicy. The website notes that candidates may use screen reading software and/or a scribe to complete the required documents, without advance notice, as all candidates are required to attest that they are the sole author of their submission. Other alternative arrangements must be requested, with supporting documentation submitted along with the request. Supporting documentation must provide evidence of review by a qualified professional, must recommend alternative arrangements that are clearly related to the identified disabilities or learning needs, and must be current.
4.6 When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.	5	In the "edTPA technical specifications" document, the model sponsor describes that a nationally representative group of teachers completed a Job Analysis Survey (JAS), which undergirds the construct validity of the edTPA. In addition, in the Transition Plan (p. 441), the bias and sensitivity review was achieved through the structured examination of handbook prompts, rubrics, and directions by a diverse and trained pool of professional teachers and teacher educators from across the nation who provided feedback on the structure of prompts, phrasing of questions, language of rubrics, and formatting of handbooks to ensure comprehensibility and equitable access and evaluation for all candidates completing edTPA. The Bias Review Meeting Orientation Manual outlines the review procedures and guidelines. The Global Handbook Bias Report presents comments and suggestions from the Bias Review Committee's review of the Assessment Handbook. Comments and suggestions generally called for review for clarity, consistency, complexity, or completion, and were tagged by bias type (e.g., language, fairness, inclusion, content, stereotype). Also, on p. 14 of the 2013 edTPA Field Test: Summary Report states that the Bias Review Committee included 10 nationally representative educators and teacher educators who reviewed draft edTPA materials for any potential bias and provided input for revision. Because edTPA is used across the nation, the nationally representative sample of reviewers is appropriate, although care should be taken to ensure that a diverse, representative sample of California educators are included in such reviews to ensure that edTPA is appropriate for California educators. The independent, external evaluators conducting this comparability investigation, all educational researchers with PhDs and five or more years of experience conducting validity studies of educational assessments, meet the intent of this Standard. The reviewers for Activity 1, however, do not represent a demographically diverse group (two females and one male Caucasian).

(continued)

Table 1.6. (Continued)

Test Design Standards from the *Joint Standards*	edTPA Rating	edTPA Rationale
4.7 The procedures used to develop, review, and try out items and to select items from the item pool should be documented.	5	The 2013 edTPA Field Test: Summary Report provided a general introduction of the assessment development, review, and field-testing procedures. For example, the model sponsor described the assessment design and architecture (p.9-12) and the item development and professional review processes (p.13-14). According to the Orientation Manual for the Bias Review Meeting, edTPA was authored and developed by a team of researchers at Stanford University, "with substantive advice from teacher educators" (p.3). The edTPA technical specifications describe the process whereby multiple groups of teachers were convened to identify the key teaching KSAs and validate that the edTPA rubrics were strongly linked to these KSAs.
4.8 The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.	4	As described in the 2013 field test summary report, the test review process included empirical analyses and the use of expert judges to review items and scoring criteria. The Orientation Manual for the Bias Review Meeting provides some details about the instructions and training that item reviewers receive; however, information on the qualifications, relevant experiences, and demographic characteristics of the judges was not found.
4.9 When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.	4	In the 2013 field test summary report, information on the field test sample size (n=18,436) and score distribution is provided (p. 437 of the edTPA Transition Plan). Of the 18,436 candidates, 3% were from California. Information on the procedures used to select the sample were not provided. Differences by demographic groups were reported (p. 272 of Transition Plan). The edTPA Transition Plan Update (May 2018) provides some details about field testing of an additional Elementary Literacy task to be combined with the Elementary Mathematics handbook. Included in the details about this field testing are the participating programs, the total number of submissions, and information about the scoring of the submissions. However, it is unclear whether the final sample was representative of the population.

(continued)

Table 1.6. (Continued)

Test Design Standards from the *Joint Standards*	edTPA Rating	edTPA Rationale
4.10 When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.	5	A polytomous item response theory (IRT) model was fit to the data to examine the theoretical foundation that underlies the use of edTPA total scores as a representation of a common construct of teaching effectiveness; the rubric levels were distributed in the expected pattern of difficulty. Evidence of model fit is provided (see p.290-291 of the Transition Plan). The unidimensional Partial Credit Model was fit to the 2014 sample of 18,436 candidates representing 17 states, including 3% from California. Models were estimated using marginal maximum likelihood as carried out with the "TAM" package in R. To evaluate fit, INFIT mean square statistics were computed for each rubric and examined to identify rubrics with INFIT values less than 0.75 or greater than 1.33, which would suggest a lack of fit. All rubric INFIT mean square statistics were within the range 0.90 to 1.15 (mean = 1.00), suggesting appropriate model-data fit for the rubrics, which was also supported by the plots of observed vs. expected scores. Differential item analyses were run to examine systematic differences in rubric difficulty for candidates with the same total scores. Analyses by major examinee groups were reported (see pgs. 63 and 272 of Transition Plan).
4.12 Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.	5	The table on p.115-150 of the Transition Plan demonstrates that the content domain of a test represented the domain defined in the test specifications. In addition, in the "edTPA technical specifications" document, the model sponsor indicated that there were pieces of evidence to support the test specifications and evidence of the validity of edTPA score interpretations was provided by the Job Analysis Survey (JAS).
4.13 When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.	4	The edTPA Transition Plan includes a section on "Analyses of Differential Effects" (see pgs. 63-64). On page 64, the model sponsor stated that, "edTPA is committed to providing an equitable assessment that is free of bias and adverse impact. While caution must be taken in making generalizations based on such small sample sizes, these findings are consistent with other reported portfolio-based performance assessment data (e.g., NBPTS, PACT, ProTeach). As more data become available, additional research is planned at the state and national levels – we are committed to supporting research to better understand these differences in performance." This statement was based on analyses in the 2015 edTPA Annual Administrative Report. No new analyses based on more recent data have been provided.

(continued)

Table 1.6. (Continued)

Test Design Standards from the *Joint Standards*	edTPA Rating	edTPA Rationale
4.15 The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.	5	Directions for test administration are presented in the Transition Plan and the edTPA website. Administration procedures are established for edTPA that accommodate candidates requiring alternative arrangements. In addition, the documents such as the edTPA Making Good Choices and the edTPA Handbooks provide clear directions for both candidates and Educator Preparation Program faculty. In the guidelines for supporting candidates document, the model sponsor listed detailed, acceptable and unacceptable forms of support for candidates within the edTPA process, which provided thorough guidelines for faculty so that they can effectively assist candidates to prepare for the assessment.
4.16 The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.	5	All these are provided in the Handbooks (e.g., see 463-517 of the Transition Plan). For each of the three tasks (planning, instruction and engaging, assessing), there are detailed introductions of the task that include "what to do," "what to submit," and "evaluation rubrics."
4.18 See Chapter 4		
4.20 See Chapter 4		
4.22 See Chapter 4		

(continued)

Table 1.6. (Continued)

Test Design Standards from the *Joint Standards*	edTPA Rating	edTPA Rationale
4.24 Test specifications should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may reduce the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.	5	The Transition Plan provides detailed evidence of the revisions to address the revised TPEs and ADS.
4.25 When tests are revised, users should be informed of the changes to the specifications, of any adjustments made to the score scale, and of the degree of comparability of scores from the original and revised tests. Tests should be labeled as “revised” only when the test specifications have been updated in significant ways.	NA	The edTPA model developer indicated that at this time, they have no changes that will require this communication.

CalTPA. Table 1.7 presents the ratings for CalTPA on each test design and development Standard from the *Joint Standards*. In Year 1 (2017–18), the technical documentation for CalTPA was limited given its field test status, and, thus, some of the Standards could not be rated in Year 1 due to unavailable technical documentation. Since that time and subsequent to the first operational administration of the revised CalTPA, the technical documentation has expanded, allowing all of the applicable test design and development *Joint Standards* to be rated in Year 2 (2018–19). The average rating across the *Joint Standards* also increased from 4.50 in Year 1 to 4.71 in Year 2. All of the Standards were rated a '4' (Evidence in the documentation mostly covers all aspects of the Standard/element) or '5' (Evidence in the documentation fully covers all aspects of the Standard/element), with the majority receiving a rating of a '5.'

Table 1.7. Ratings on the Joint Standards for CalTPA

Joint Standard	CalTPA Rating	CalTPA Rationale
4.1 Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	5	Per the CalTPA Assessment Overview document (available on the CalTPA website), the redeveloped CalTPA is intended to provide both a formal assessment of candidate ability and a framework of performance-based guidance during the candidate's teacher preparation program to inform candidate preparation and continued professional growth through induction. Performance data is shared with institutions to assist them in making program improvements and to guide induction programs as they work with new teachers to individualize learning plans. The domains measured are outlined in the CalTPA to TPE Map. The Rules of Participation presented on the CalTPA Policies webpage state that CalTPA is only to be taken by individuals fulfilling a program requirement and/or a California preliminary teaching credentialing requirement. There is evidence of face validity and content validity through the nature of the performance tasks and the clear link to the TPEs. There is evidence of construct validity from correlational analysis and factor analysis conducted using field test data. The results from these analyses provide evidence that the rubrics represent a common underlying construct and also support the organization of the rubrics within steps within the two test cycles.
4.2 In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.	5	Though not listed in a single specifications document, the Assessment Guides (including scoring rubrics), CalTPA to TPE Map, and CalTPA website in combination describe the intended uses of the test, the test length (two instructional cycles, each containing several steps), the item formats (performance tasks). The two cycles are ordered so that the first cycle will inform the second. Candidates complete the assessment during field placements as part of their teacher preparation program. Directions for completing the assessment are available in the Assessment Guides. Submission guidelines are presented on the model website, along with links to available supports. The accommodations policy is outlined on the model website. Materials to be created/completed/submitted are described in the Assessment Guide. Results from correlational and factor analyses provide evidence that the rubrics represent a common underlying construct and also support the organization of the rubrics within steps within the two test cycles.

(continued)

Table 1.7. (Continued)

Joint Standard	CalTPA Rating	CalTPA Rationale
4.2 (continued)		Pearson's ePEN application helps CalTPA scoring leads monitor the means, standard deviations, and interrater reliability of assessors over time, by assessor, subject area, rubric, etc. Scoring leads members are trained to monitor for scoring drift and other issues via automatically produced reports in the system. During operational scoring, CalTPA collects and monitors double scoring inter-rater reliability and validity (reliability) scoring. The descriptive statistics for CalTPA (by rubric and cycle and by credential and by demographic groups) are computed (2018–19 data provided in appendix to Standard Setting documentation). CalTPA assessor qualifications were outlined in the PowerPoint presentation from the CalTPA Design Team meeting (July 2018). These included being a current or recently retired California education professional with specific content expertise. Scorer training, scoring, and reporting for the field test cycle are also described in the Update on the Redevelopment of the California Teaching Performance Assessment.
4.5 If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	5	The CalTPA Alternative Arrangements webpage outlines the process for requesting alternative conditions. According to the policy, it is acceptable for all candidates to use screen reading software and/or a scribe to complete CalTPA submissions as long as the content of the submission is the original work of the candidate. The rationale for this is offered. Extensive documentation is required for alternative arrangement due to diagnosed disability. Alternative arrangements may also be requested regarding the provision of video evidence. Settings in which a candidate may be teaching but which are not appropriate for video recording (e.g., a juvenile correctional facility), are grounds for using audio recording or written transcription. This arrangement requires a detailed description of the procedures for submission.
4.6 When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.	4	A Design Team made up of experts external to the Commission was assembled and serves in an advisory capacity for all aspects of assessment design. The Design Team includes "twenty-one members representing the full range of teacher preparation programs, teacher induction programs, and the geographic regions of California."

(continued)

Table 1.7. (Continued)

Joint Standard	CalTPA Rating	CalTPA Rationale
4.6 (continued)		External content expert panels were convened in October 2016 to review the CalTPA Instructional Cycles, rubrics, and materials for subject-specific appropriateness. Panel members included teachers, TK-12 district-level staff, university faculty, and Department of Education staff from throughout the state, though relevant experiences and other demographic information was not reported. The CalTPA Bias Review Committee convened in August 2017 and again in August 2018 to review the redeveloped CalTPA and comment on potential bias issues. Background and demographic characteristics of those committee members (including whether these experts are "external to the testing program") were not included in the documentation. The independent, external evaluators conducting this comparability investigation, all educational researchers with PhDs and five or more years of experience conducting validity studies of educational assessments, meet the intent of this Standard. The reviewers for Activity 1, however, do not represent a demographically diverse group (two females and one male Caucasian).
4.7 The procedures used to develop, review, and try out items and to select items from the item pool should be documented.	5	Materials from all monthly Design Team meetings are maintained by the Commission and made available for this documentation review. Their major activities during the development and pilot phases are also summarized in the Update on the Redevelopment of the CalTPA document. Field Testing occurred using a broad sample of programs and candidates and field test data was used to adjust instructions, prompts, and rubrics. Surveys were administered to candidates, program coordinators, and assessors. Those survey results were used to identify areas for improvement.
4.8 The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.	4	Both empirical analyses and the use of expert judges were used to review items and scoring criteria. Instructions and training for expert judges are provided in the Bias Review Conference instructions (CalAPA-TPA_OP_BRC_Instructions.pdf) and the Bias Review Conference PowerPoint slides (CalTPA_2018 BRC_Orientation.ppt). Other than geographic area and professional affiliation, background information on the Bias Review Committee was not included in the documentation.

(continued)

Table 1.7. (Continued)

Joint Standard	CalTPA Rating	CalTPA Rationale
4.9 When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.	5	The redesigned CalTPA was piloted between January and April of 2017. The pilot sample is described in the Update on the Redevelopment of the CalTPA document. The final pilot sample included 250 candidates from programs throughout the state, approximately 67% female and 34% non-white, representing a range of content areas, program types, program lengths, field placement types and field placement settings. The CalTPA was field tested from October 2017 through April 2018.The criteria for the selection of institutions to participate in the Field Test is presented in the Update on the Redevelopment of the California Teaching Performance Assessment document. The field test collected data from approximately 900 candidates across a sample of institutions that reflected the diversity of program types, sizes, and candidates served by institutions, and service areas in California. The criteria for selecting this diverse sample are described in the aforementioned document; content area representation is provided in Table 3 of the aforementioned document.
4.10 When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.	4	The field test sample on which psychometric analyses are based is described in the Update on the Redevelopment of the California Teaching Performance Assessment and in the PowerPoint presentation from the CalTPA Design Team meeting (July 2018). There were approximately 900 candidates in the field test sample, and the sample was representatively diverse. Model based methods are not used to estimate item parameters for the CalTPA. Overall scores are pass/fail decisions. The Update on the Redevelopment of the California Teaching Performance Assessment explains that field test findings were used to revise and reduce the number of rubrics. However, the process by which the performance data was used to aid these decisions (e.g., screening components of the assessment or particular rubrics for difficulty, discrimination, or differential functioning for major examinee groups) is not documented.
4.12 Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.	5	The CalTPA to TPE Map documents the extent to which the content domain of the assessment represents the TPEs, per the model sponsors' determinations.

(continued)

Table 1.7. (Continued)

Joint Standard	CalTPA Rating	CalTPA Rationale
4.13 When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.	5	Potential sources of irrelevant variance include measurement error that is introduced during the scoring process or systematic differences in the observed performances of subgroups that are not related to true differences in their mastery of the TPEs. The potential for irrelevant variance introduced during scoring was addressed through scorer training, qualification, and calibration. The Update on the Redevelopment of the California Teaching Performance Assessment and the PowerPoint presentation from the CalTPA Design Team meeting (July 2018) provide details about the required scorer qualifications and the process scorers had to successfully complete to be qualified to score operationally. HumRRO staff also observed scoring and noted that scorers were trained on the range of scores and were required to reach exact agreement on a minimum number of rubrics to meet calibration requirements. Irrelevant variance related to subgroup differences was addressed during the development phase though item reviews. The Update on the Redevelopment of the California Teaching Performance Assessment and the PowerPoint presentation from the CalTPA Design Team meeting (July 2018) describe how the Design Team and other review committees combined both their expertise and feedback from the field to continuously revise the tasks, directions and rubrics. Additional bias reviews were conducted by the model sponsor in August 2018.
4.15 The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.	5	The evidence to be submitted is clearly described in the Assessment Guides. The submission guidelines are presented on the CalTPA website, as are instructions for requesting alternative arrangements, including arrangements for alternatives to video recording. Furthermore, the model sponsors provided numerous implementation presentations to aid understanding of implementing/administering CalTPA.

(continued)

Table 1.7. (Continued)

Joint Standard	CalTPA Rating	CalTPA Rationale
4.16 The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.	5	All steps of the instructional cycles and the associated rubrics are provided to candidates, along with specific questions to be answered and evidence to be submitted. Based on qualitative information obtained from the CalTPA field test, exemplars of responses to cycles were developed.
4.18 See Chapter 4		
4.20 See Chapter 4		
4.22 See Chapter 4		
4.24 Test specifications should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may reduce the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.	5	As demonstrated in the Update on the Redevelopment of the CalTPA and the Faculty Resources page of the model's website, CalTPA is undergoing several changes in response to changes in the TPEs and the Assessment Design Standards.
4.25 When tests are revised, users should be informed of the changes to the specifications, of any adjustments made to the score scale, and of the degree of comparability of scores from the original and revised tests. Tests should be labeled as "revised" only when the test specifications have been updated in significant ways.	4	Information about planned and ongoing changes to the CalTPA are available to users via the model's website and via presentations provided by model sponsors. The model sponsors appropriately identify the current version of the CalTPA as the "Redeveloped CalTPA," in light of significant changes that have been made to address revisions to the TPEs and ADS. Information on the degree of comparability of scores from original and revised tests is not available.

Comparison of the Strength of Evidence for Assessment Design Standards across TPA Models

Table 1.8 summarizes the ratings on each *Assessment Design Standard* for all three models (sans the rationale for each rating). The ratings for all three models are similar. The average ratings were 4.83, 4.76, and 4.83 for FAST, edTPA and CalTPA, respectively. Most ADS received ratings of '5' (Evidence in the documentation fully covers all aspects of the Standard/element) across all three models. None of the models received any ratings below '4' (Evidence in the documentation mostly covers all aspects of the Standard/element).

Each model received a rating of '4' on three or four Standards; three of those Standards were common to all three models, which are: ADS 1(i), ADS 1(l) and ADS 2(f). First, ADS 1(i) states that, "*The model sponsor provides a clear statement acknowledging the intended uses of the assessment. The statement demonstrates the model sponsor's clear understanding of the implications of the assessment for candidates, preparation programs,* ***the public schools, and TK-12 students*** [emphasis added]." While all three models provided clear statements acknowledging the intended uses of the assessment for candidates and preparation programs, none of the models included clear statements on the implications of the assessment for public schools and TK-12 students in the documents and materials that were reviewed. The models could attain a rating of '5' (fully covers all aspects of the Standard/element) if they were to include clear statements about the implications of the assessment for public schools and TK-12 students in their documentation for their model. Second, ADS 1(l) states that, "*In designing assessment administration procedures, the model sponsor includes administrative accommodations that preserves assessment validity while addressing issues of access for candidates with disabilities or learning needs*." All three models provide clear guidance on procedures for requesting accommodations; however, there is no empirical evidence demonstrating the comparability of scores on accommodated assessments. If this evidence were to be provided, all models would attain a '5' on ADS 1(l). Finally, ADS 2(f) states that, "*Model sponsors must document that all candidate appeals granted a second scoring are scored by a new assessor unfamiliar with the candidate or candidate's response*." While procedure documents from all three models indicate that this is part of their procedures, there was no documentation provided as evidence to document that this occurred. Should such documentation be provided, then models would attain a '5' on ADS 2(f).

Table 1.8. Comparison of Ratings on Assessment Design Standards across TPA Models

Assessment Design Standard	FAST Rating	edTPA Rating	CalTPA Rating
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 rubrics that are clearly related to the TPEs that the task measures. Each task and its associated rubrics measure two or more TPEs. Collectively, the tasks and rubrics in the assessment address key aspects of the six major domains of the TPEs. The sponsor of the performance assessment documents the relationships between TPEs, tasks and rubrics.	See Chpt. 4	See Chpt. 4	See Chpt. 4
1(b) The TPA model sponsor must include a focus on content-specific pedagogy within the design of the TPA tasks and scoring scales to assess the candidate's ability to effectively teach the content area(s) authorized by the credential.	5	5	5
1(c) Consistent with the language of the TPEs, the model sponsor defines scoring rubrics so candidates for credentials can earn acceptable scores on the Teaching Performance Assessment with the use of different content-specific pedagogical practices that support implementation of the TK-12 content standards and curriculum frameworks. The model sponsor takes steps to plan and anticipate the appropriate scoring of candidates who use a wide range of pedagogical practices that are educationally effective and builds scoring protocols to take these variations into account.	See Chpt. 4	See Chpt. 4	See Chpt. 4
1(d) The model sponsor must include within the design of the TPA candidate tasks a focus on addressing the teaching of English learners, all underserved education groups or groups that need to be served differently, and students with special needs in the general education classroom to adequately assess the candidate's ability to effectively teach all students.	5	5	5
1(e) For Multiple Subject candidates, the model sponsor must include assessments of the core content areas of at least Literacy and Mathematics. Programs use local program performance assessments for History-Social Science and Science if not already included as part of the TPA.	5	5	5
1(f) The model sponsor must include a focus on classroom teaching performance within the TPA, including a video of the candidate's classroom teaching performance with candidate commentary describing the lesson plan and rationale for teaching decisions shown and evidence of the effect of that teaching on student learning.	5	5	5
1 (g) The TPA model sponsor must provide materials appropriate for use by programs in helping faculty become familiar with the design of the TPA model, the candidate tasks and the scoring rubrics so that faculty can effectively assist candidates to prepare for the assessment. The TPA model sponsor must also provide candidate materials to assist candidates in understanding the nature of the assessment, the specific assessment tasks, the scoring rubrics, submission processes and scoring processes.	See Chpt. 4	See Chpt. 4	See Chpt. 4
1(h) The model sponsor develops scoring rubrics 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	See Chpt. 4	See Chpt. 4	See Chpt. 4

(continued)

Table 1.8. (Continued)

Assessment Design Standard	FAST Rating	edTPA Rating	CalTPA Rating
1(i) The model sponsor provides a clear statement acknowledging the intended uses of the assessment. The statement demonstrates the model sponsor's clear understanding of the implications of the assessment for candidates, preparation programs, the public schools, and TK-12 students. The statement includes appropriate cautions about additional or alternative uses for which the assessment is not valid. All elements of assessment design and development are consistent with the intended uses of the assessment for determining the pedagogical competence of candidates for Preliminary Teaching Credentials in California and as information useful for determining program quality and effectiveness.	4	4	4
1(j) The model 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.	5	5	5
1(k) The model sponsor completes initial and periodic basic psychometric analyses to identify pedagogical assessment tasks and/or scoring rubrics that show differential effects in relation to candidates' race, ethnicity, language, gender or disability. When group pass-rate differences are found, the model sponsor investigates the potential sources of differential performance and seeks to eliminate construct-irrelevant sources of variance.	5	4	5
1(l) In designing assessment administration procedures, the model sponsor includes administrative accommodations that preserve assessment validity while addressing issues of access for candidates with disabilities or learning needs.	4	4	4
1(m) In the course of determining a passing standard, the model sponsor secures and reflects on the considered judgments of teachers, supervisors of teachers, 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 model sponsor periodically reviews the reasonableness of the scoring scales and established passing standard, when and as directed by the Commission.	See Chpt. 5	See Chpt. 5	See Chpt. 5
1(n) To preserve the validity and fairness of the assessment over time, the model sponsor may need to develop and field test new pedagogical assessment tasks and multi-level scoring rubrics to replace or strengthen prior ones. Initially and periodically, the model sponsor analyzes the assessment tasks and scoring rubrics to ensure that they yield important evidence that represents candidate knowledge and skill related to the TPEs, and serve as a basis for determining entry-level pedagogical competence to teach the curriculum and student population of California's TK-12 public schools. The model sponsor documents the basis and results of each analysis, and modifies the tasks and rubrics as needed.	5	5	5
1(o) The model sponsor must make all TPA materials available to the Commission upon request for review and approval, including materials that are proprietary to the model sponsor. The Commission will maintain the confidentiality of all materials designated as proprietary by the model sponsor.	5	5	5
2(a) In relation to the key aspects of the major domains of the TPEs, the pedagogical assessment tasks, rubrics, and the associated directions to candidates are designed to yield enough valid evidence for an overall judgment of each candidate's pedagogical qualifications for a Preliminary Teaching Credential as one part of the requirements for the credential.	See Chpt. 4	See Chpt. 4	See Chpt. 4
2(b) Pedagogical assessment tasks and scoring rubrics are extensively field tested in practice before being used operationally in the Teaching Performance Assessment. The model sponsor evaluates the field test results thoroughly and documents the field test design, participation, methods, results and interpretation.	5	5	5

(continued)

Table 1.8. (Continued)

Assessment Design Standard	FAST Rating	edTPA Rating	CalTPA Rating
2(c) The Teaching Performance Assessment system includes a comprehensive process to select and train assessors who score candidate responses to the pedagogical assessment tasks. An assessor training program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and the multi-level scoring rubrics. The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy and calibration in relation to the scoring rubrics associated with the task. The model sponsor establishes selection criteria for assessors of candidate responses to the TPA. The selection criteria include but are not limited to appropriate pedagogical expertise in the content areas assessed within the TPA. The model sponsor selects assessors who meet the established selection criteria and uses only assessors who successfully calibrate during the required TPA model assessor training sequence. When new pedagogical tasks and scoring rubrics are incorporated into the assessment, the model sponsor provides additional training to the assessors, as needed.	See Chpt. 4	See Chpt. 4	See Chpt. 4
2(d) In conjunction with the provisions of the applicable Teacher Preparation Program Standards relating to the Teaching Performance Assessment, the model 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.	5	5	5
2(e) The model sponsor provides a consistent scoring process for all programs using that model, including programs using a local scoring option provided by the model sponsor. The scoring process conducted by the model sponsor to assure the reliability and validity of candidate outcomes on the assessment may include, for example, regular auditing, selective back reading, and double scoring of candidate responses near the cut score by the qualified, calibrated scorers trained by the model sponsor. All approved models must include a local scoring option in which the assessors of candidate responses are program faculty and/or other individuals identified by the program who meet the model sponsor's assessor selection criteria. These local assessors are trained and calibrated by the model sponsor, and whose scoring work is facilitated and their scoring results are facilitated and reviewed by the model sponsor. The model sponsor provides a detailed plan for establishing and maintaining scorer accuracy and inter-rater reliability during field testing and operational administration of the assessment. The model 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. The model sponsor must provide an annual audit process that documents that local scoring outcomes are consistent and reliable within the model for candidates across the range of programs using local scoring, and informs the Commission where inconsistencies in local scoring outcomes are identified. If inconsistencies are identified, the sponsor must provide a plan to the CTC for how it will address and resolve the scoring inconsistencies both for the current scoring results and for future scoring of the TPA.	See Chpt. 4	See Chpt. 4	See Chpt. 4

(continued)

Table 1.8. (Continued)

Assessment Design Standard	FAST Rating	edTPA Rating	CalTPA Rating
2(f) The model sponsor's assessment design includes a clear and easy to implement 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, if the program is using centralized scoring provided by the model sponsor. If the program is implementing a local scoring option, the program must provide an appeal process as described above for candidates who do not pass the assessment. Model sponsors must document that all candidate appeals granted a second scoring are scored by a new assessor unfamiliar with the candidate or the candidate's response.	4	4	4
2(g) The model sponsor conducting scoring for the program provides results on the TPA to the individual candidate based on performance relative to TPE domains and/or to the specific scoring rubrics within a maximum of three weeks following candidate submission of completed TPA responses. The model sponsor provides results to programs based on both individual and aggregated data relating to candidate performance relative to the rubrics and/or domains of the TPEs. The model sponsor also follows the timelines established with programs using a local scoring option for providing scoring results.	See Chpt. 4	See Chpt. 4	See Chpt. 4
2(h) The model sponsor provides program level aggregate results to the Commission, in a manner, format and time frame specified by the Commission, as one means of assessing program quality. It is expected that these results will be used within the Commission's ongoing accreditation system.	5	5	5
3(a) The model sponsor provides technical assistance to programs implementing the model to support fidelity of implementation of the model as designed. Clear implementation procedures and materials such as a candidate and a program handbook are provided by the model sponsor to programs using the model.	5	5	5
3(b) A model sponsor conducting scoring for programs is responsible for providing TPA outcomes data at the candidate and program level to the program within three weeks and to the Commission, as specified by the Commission. The model sponsor supervising/moderating local program scoring oversees data collection, data review with programs, and reporting.	5	5	5
3(c) The model sponsor is responsible for submitting at minimum an annual report to the Commission describing, among other data points, the programs served by the model, the number of candidate submissions scored, the date(s) when responses were received for scoring, the date(s) when the results of the scoring were provided to the preparation programs, the number of candidate appeals, first time passing rates, candidate completion passing rates, and other operational details as specified by the Commission.	NA[a]	NA	NA
3(d) The model sponsor is responsible for maintaining the currency of the TPA model, including making appropriate changes to the assessment tasks and/or to the scoring rubrics and associated program, candidate, and scoring materials, as directed by the Commission when necessitated by changes in TK-12 standards and/or in teacher preparation standards.	5	5	5
3(e) The model sponsor must define the retake policies for candidates who fail one or more parts of the TPA which preserve the reliability and validity of the assessment results. The retake policies must include whether the task(s) on which the candidate was not successful must be retaken in whole or in part, with appropriate guidance for programs and candidates about which task and/or task components must be resubmitted for scoring by a second assessor and what the resubmitted response must include.	5	5	5
Average	4.83	4.76	4.83

Note. NA = Not applicable.

[a]At the time this investigation was being conducted, the Commission was not requiring an annual report.

Comparison of the Strength of Evidence for the Joint Standards across TPA Models

Table 1.9 presents the ratings on each test design and development Standard from the *Joint Standards* for all three models (sans the rationale for each rating). The ratings for all three models are similar. The average ratings were 4.64, 4.77, and 4.71 for FAST, edTPA and CalTPA, respectively. Most ADS received ratings of '5' (Evidence in the documentation fully covers all aspects of the Standard/element) across all three models. None of the models received a rating below a '4' (Evidence in the documentation mostly covers all aspects of the Standard/element).

Each model received a rating of '4' on three to five Standards; one of those Standards was common to all three models: JS 4.8. This Standard states that, "*The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive*." The FAST model did not provide details on the instructions and training given to reviewers and the edTPA and CalTPA models did not provide details on the qualifications, relevant experiences, and demographic characteristics of the judges. Consequently, each model received a rating of '4' as opposed to '5' on this Standard.

Table 1.9. Comparison of Ratings on the Joint Standards across TPA Models

Test Design Standards from the *Joint Standards*	FAST Rating	edTPA Rating	CalTPA Rating
4.1 Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	5	5	5
4.2 In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.	4	5	5
4.5 If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	5	5	5
4.6 When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.	4	5	4
4.7 The procedures used to develop, review, and try out items and to select items from the item pool should be documented.	5	5	5

(continued)

Table 1.9. (Continued)

Test Design Standards from the *Joint Standards*	FAST Rating	edTPA Rating	CalTPA Rating
4.8 The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.	4	4	4
4.9 When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.	5	4	5
4.10 When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.	4	5	4
4.12 Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.	5	5	5
4.13 When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.	5	4	5
4.15 The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.	5	5	5
4.16 The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.	5	5	5
4.18 Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.	See Chpt. 4	See Chpt. 4	See Chpt. 4
4.20 The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.	See Chpt. 4	See Chpt. 4	See Chpt. 4

(continued)

Table 1.9. (Continued)

Test Design Standards from the *Joint Standards*	FAST Rating	edTPA Rating	CalTPA Rating
4.22 Test developers should specify the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.	See Chpt 4	See Chpt 4	See Chpt 4
4.24 Test specifications should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may reduce the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.	5	5	5
4.25 When tests are revised, users should be informed of the changes to the specifications, of any adjustments made to the score scale, and of the degree of comparability of scores from the original and revised tests. Tests should be labeled as "revised" only when the test specifications have been updated in significant ways.	4	NA	4
Average	4.64	4.77	4.71

Note. NA = Not applicable.

Discussion

Activity 1 served as an overarching investigation, via a documentation review, of the eight claims identified in the introduction of this report. This activity involved a comprehensive review and comparison of the documents and materials developed by each model sponsor. The evidence was reviewed, and evaluations were made regarding the strength of evidence for adherence to the ADS, which are reflective of the eight claims. We also reviewed and evaluated the documentation for adherence to the test design and development Standards from the *Joint Standards.* The latter evaluation was conducted to ensure that the documentation and materials for each TPA model also cover industry-wide principles for test design and development.

In Year 1 of this comparability investigation, the available technical documentation for FAST and CalTPA, which were both being field tested, was limited and sparse. In Year 2, additional and more detailed documentation became available for FAST and CalTPA. Moreover, additional detail and clarification on the available documentation was provided for all three models in Year 2. As a result, the average ratings for adherence to Standards (both the ADS and *Joint Standards*) increased from Year 1 to Year 2, particularly for FAST and CalTPA, which had comparatively lower ratings than edTPA in Year 1. As such, the technical documentation indicates that all three TPA models mostly or fully adhere to the ADS and *Joint Standards.*

The reader is referred to Table 1.9 for cases in which individual models could be strengthened. All three models could further strengthen adherence to the Standards through the following:

- Including in their documentation statements acknowledging the intended uses of the assessment to public schools and TK-12 students, both of which are requirements of ADS 1(i).
- Collecting empirical evidence of the comparability of scores on accommodated assessments [ADS 1(l)].

- Providing documentation that candidate appeals granted a second scoring were indeed scored by a new assessor unfamiliar with the candidate or candidate's response [ADS 2(f)].
- Providing additional detail on the qualifications, relevant experiences, and demographic characteristics of expert judges used to review the assessments (edTPA and FAST) and on the instructions and training provided to those expert judges (FAST) [JS 4.8].

Conclusion

Overall, the findings from Activity 1 indicate that the TPA models either mostly or fully adhere to the ADS and *Joint Standards* with regard to the documentation and evidence provided for each model. This provides support for Claim 1, which states in part that, "*The TPA models are sufficiently comparable in their representation of the Commission's Assessment Design Standards*."

Chapter 2: Content Validity Comparability Analysis (Activity 2)

Andrea L. Sinclair & Arthur Thacker

Introduction

Activity 2 builds upon Activity 1 by further investigating Claim 1:

> The TPA models are sufficiently comparable in their representation of the Commission's *Assessment Design Standards and* in their assessment and weighting of the Commission-adopted *Teaching Performance Expectations* (TPEs).

Whereas Activity 1 focused on the *Assessment Design Standards* (ADS), this activity focused on the depth and breadth with which each TPA model assesses each TPE element. Activity 2 is essentially an external content validity investigation to ensure that each TPA model does, in fact, measure the targeted knowledge, skills, and abilities (KSAs) specified by the TPEs. This is a prerequisite for ensuring that the three TPA models are comparable.

Assessment Design Standard 1(a) states in part that, "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. … Collectively, the tasks and rubrics in the assessment address key aspects of the six major domains of the TPEs."[13] Thus, for this activity, in Year 1 of the comparability study, two external panels of teacher preparation experts—hereafter referred to as subject matter experts (SMEs)—were convened at a four-day workshop to provide expert evaluations of the coverage of the TPEs by each TPA model.[14] Because the TPA models are complex—each consisting of multiple components (i.e., edTPA Tasks 1–3, CalTPA Cycles 1 and 2, FAST Site Visitation Project and Teaching Sample Project) and requiring multiple forms of evidence (e.g., lesson plans, videos, self-reflections)—the SMEs' evaluations extended beyond a simple evaluation of coverage of TPE elements by TPAs. The SMEs also identified the type of evidence required by each model to assess each TPE element and an evaluation of how thoroughly each model assesses the KSAs specified by each TPE element. At the end of the workshop, the two panels of SMEs produced a cross-validated matrix (one for each TPA model) mapping each TPE element to each component of the TPA according to the type of evidence required by the TPA component (e.g., lesson plan, video, reflection), along with an evaluation of how thoroughly the collective set of evidence required by the TPA assesses the KSAs in each TPE element. These matrixes allow for comparisons across the TPA models with regard to the sufficiency of evidence for the TPEs, and, thus, allow for conclusions regarding comparability of the content validity of the TPA models.

The detailed method and results from the Year 1 content validity workshop can be found in the Year 1 report (Sinclair & Thacker, 2018). Following Year 1 of the comparability study (2017–18), changes were made to the models, most notably to FAST and CalTPA, which were field tested in 2017–18 and operational in 2018–19. The FAST model revised and clarified candidate instructions and the wording of the rubrics. The CalTPA model also revised some wording in its assessment guides and rubrics, but also made other substantive changes, such as reducing the number of rubrics and revising scorer training procedures. The edTPA implemented an

[13] See page 59 in the Method section of this chapter for an explanation of how "key aspects" was defined for this activity.

[14] The teacher preparation experts were "external" in the sense that they were not involved in the design or revision of any of the TPA models.

additional elementary education handbook in 2018–19 (which had been field tested in 2017–18), but the edTPA otherwise remained unchanged from 2017–18 to 2018–19.

Given the changes to the models following the content validity workshop in April 2018, we revisited the TPE-to-TPA mappings that stemmed from that workshop. In discussion with the Technical Advisory Committee (TAC), the decision was made to convene a subset of the 14 SMEs who participated in the April 2018 workshop and have them review the updates to the TPA models and then make changes to the TPE-to-TPA mappings, as warranted.

In the sections that follow, we present the methodology, results, discussion and conclusions for this activity.

Method

In July 2019, the SMEs who participated in the April 2018 workshop were individually contacted via email to determine their availability to participate in a series of webinars on August 14 – 16th, 2019.[15] Five SMEs from the April 2018 workshop were identified for participation. Four of the SMEs had experience as teacher preparation experts with the revised CalTPA, three with the edTPA, and one with the revised FAST. All but one of the SMEs had 10+ years of experience as teacher preparation experts.

Prior to the webinar, example portfolios from 2018–19 and the corresponding manual/handbook/guide for each model were securely shared with the SMEs.[16] Model representatives from each TPA model were invited to the first hour of the first day of the webinar to provide an overview of the changes implemented in 2018–19. The FAST model sponsor was not available for the webinar; thus, a crosswalk between the 2017–18 rubrics and 2018–19 rubrics was shared with SMEs to demonstrate the changes to the FAST rubric. The SME from Fresno State was also able to elaborate on changes to FAST that were implemented in 2018–19.

Following the discussion of changes implemented in 2018–19, the SMEs systematically reviewed each linkage of each TPE element to each TPA component. We used the TPE-to-TPA linkage matrices (i.e., the product from the Year 1 content validity workshop) as the starting point. We began with CalTPA. First, the SMEs reviewed each TPE element for Cycle 1. The question they asked is, “Does this component of CalTPA assess this TPE element, and, if so, what kind of evidence(s) must a candidate submit to demonstrate the knowledge, skills, and abilities (KSAs) required by that TPE element?” They did this for each TPE element for Cycle 1 and then repeated that same process for Cycle 2. First, they determined whether any “unlinked” TPE elements should be linked and, secondly, should any previously linked elements be linked with additional evidence requirements. After they completed this process for Cycle 1 and Cycle 2, they concluded the review by revisiting the “Strength of Evidence” ratings from the April 2018 workshop. This same process was repeated for edTPA and FAST. Strength of evidence was rated on the following scale:

[15] It was outside the scope of the current contract to conduct another onsite, four-day workshop with 14 SMEs. Thus, conducting a virtual workshop with a subset of the SMEs from the April 2018 workshop was the best option given the constraints. This approach was discussed with and approved by the TAC before proceeding.

[16] Given that the multiple-subject credential is the most frequently sought credential, the materials shared with the SMEs were for multiple subject/elementary education, although the focus of the webinar was on the TPE elements that are common across all credential areas.

- Weak evidence (red) = The model measures this TPE, but the evidence only requires a shallow demonstration of the KSAs required by this TPE (i.e., basic recall & reproduction), *and/or* omits key aspects of the TPE.
- Moderate evidence (yellow) = The model measures this TPE and the evidence requires mental processing beyond recall and reproducing; it requires comprehension and subsequent processing of information, *and* it covers key aspects of the TPE.
- Strong evidence (green) = The model measures this TPE and the evidence requires evaluation of multiple sources of information or application of significant conceptual understanding and higher-order thinking, *and* it covers the full breadth and depth of the TPE.

It should be noted that because many of the TPE elements contain several components the criterion for obtaining a "strong evidence" rating was quite stringent. Take, for example, TPE element 4.4 which states:

> Plan, design, implement and monitor instruction, making effective use of instructional time to maximize learning opportunities and provide access to the curriculum for all students by removing barriers and providing access through instructional strategies that include:
>
> - appropriate use of instructional technology, including assistive technology;
> - applying principles of UDL and MTSS;
> - use of developmentally, linguistically, and culturally appropriate learning activities, instructional materials, and resources for all students, including the full range of English learners;
> - appropriate modifications for students with disabilities in the general education classroom;
> - opportunities for students to support each other in learning; and
> - use of community resources and services as applicable.

As discussed in the consensus discussions between the two, 7-person panels during the April 2018 workshop, and which was carried over to the August 2019 webinars, if the SMEs did not find evidence that the model deeply assessed all aspects of the TPE element, then the model did not receive a "strong evidence" rating. If the model omitted any key aspects of the TPE element and/or required only a shallow demonstration of the KSAs required by the TPE element, then the model received a "weak evidence" rating for that TPE element. Through consensus discussion, panelists agreed that a "key aspect" of a TPE element was a specific activity described within the TPE element. For example, for TPE element 4.4 displayed above, the "key aspects" are: plan instruction, design instruction, implement instruction, monitor instruction, make use of instructional time to maximize learning opportunities, provide access to the curriculum for all students by removing barriers, and provide access to the curriculum for all students through the bulleted list of instructional strategies. If the TPE element used the word "include" or "including" (as with TPE element 4.4), then the text that follows had to be assessed by the model to receive a strong evidence rating. The panelists were mindful of the TPE elements that used the word "or" or "and/or." If the TPE element used "or" or "and/or," then that served as a signal that not all components of the TPE had to be assessed by the model in order for the element to receive a strong evidence rating. An example of this is TPE element 4.6 which states, "Access resources for planning and instruction, including the expertise of

community and school colleagues through in-person *or* [emphasis added] virtual collaboration, co-teaching, coaching *and/or* [emphasis added] networking.

Results

Changes that were made to the April 2018 linkage matrices following the August 2019 webinars are highlighted in yellow in Tables 2.1 to 2.6. Given that edTPA did not make any changes to the rubrics or evidence requirements in 2018–19, the SMEs ultimately decided, after reviewing the linkages from 2018, not to make any changes to the linkage matrix for edTPA.[17] Thus, there are no yellow highlights denoting changes in Tables 2.1 to 2.6 for edTPA.[18] For CalTPA and FAST, the majority of the linkages from Year 1 (April 2018) strength remained intact, although there were some notable updates.

For CalTPA, two TPE elements that were previously *un*linked became linked (TPEs 3.8 and 6.5), although the SMEs rated the of evidence for these new linkages as "weak." Regarding additional evidence requirements, across all the TPE elements, nine additional evidence requirements were linked to TPE elements in Cycle 1 and four additional evidence requirements were linked to TPE elements in Cycle 2 for a total of 13 new evidence linkages. Finally, for TPE elements that were already linked to Cycle 1 and/or Cycle 2, the strength of evidence rating increased for six of those TPE elements (two of which increased to a strong rating—TPEs 3.5 and 4.1).

For FAST, two TPE elements that were previously *un*linked became linked (TPEs 3.8 and 5.4), although the SMEs rated the strength of evidence for these new linkages as "weak." Regarding additional evidence requirements, across all the TPE elements, two additional evidence requirements were linked to TPE elements; both linkages were for the TSP component of FAST. Finally, for TPE elements that were already linked to SVP and/or TSP, the strength of evidence rating increased for two of those TPE elements (TPEs 3.2 and 3.5); in both cases, the strength of evidence increased to a strong rating.

[17] In the April 2018 content validity workshop, the SMEs mapped the Planning (Task 1), Instruction (Task 2), and Assessment (Task 3) rubrics to the TPEs. They did not map the additional Assessment Task (Task 4) for elementary education to the TPEs.

[18] We note that edTPA is a nationally validated assessment program used in 44 states. edTPA was initially approved for use in California in 2014 having illustrated alignment to the TPEs. In 2015 the Commission adopted revised Assessment Design Standards and in 2016 the Commission adopted revised TPEs. The edTPA was again approved by the Commission in 2018, having demonstrated alignment to the TPEs.

Table 2.1. TPE 1: Evidence Types Mapped to TPA Components and Strength of Evidence Dashboard

TPE 1: Engaging and Supporting all Students in Learning	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
1. Apply knowledge of students, including their prior experiences, interests, mental health and social-emotional learning needs, as well as their funds of knowledge, cultural, language, and socio-economic backgrounds to engage them in learning.	Cnt Les InM Cmt	Vid Cmt	Cnt	●		Cnt Les AsD Cmt	Cnt Cmt	●		Cnt Les Cmt	Cnt Cmt Cmp	●
2. Maintain ongoing communication with students and parents regarding achievement expectations and support needs.	--[a]	--[a]	Fdk Cmt	●[b]		--[a]	AsD Cmt	●[b]		--[a]	Asm AsD Fbk Cmt	●
3. Connect subject matter to real-life contexts and provide hands-on experiences to engage student interest, support student motivation, and allow students to extend their learning.	Les InM Asm Cmt	Vid Cmt	--[a]	●		Les Cmt	Cnt Cmt	●		Les Cmt	Les Cmt	●
4. Use a variety of developmentally and ability-appropriate instructional strategies, resources, and assistive technology, including principles of Universal Design and a Multi-tiered System of Supports (MTSS), to support access to the curriculum for a wide range of learners within the (general education) classroom (and/or learning environment).	Cnt Les InM Asm Cmt	Vid Cmt	Cnt AsD Cmt	●		Les Cmt	Les Asm Cmt	●			Les Asm AsD Cmt	●
5. Promote students' critical and creative thinking and analysis through activities that provide opportunities for inquiry, collaborative problem solving, responding to and framing meaningful questions, and reflection.	--[a]	--[a]	Cmt	●		Les Vid Cmt	Les Vid Asm Cmt	●		Les Vid	Les Cmt	●
6. Provide a supportive learning environment for students' first and/or second language acquisition by using research-based instructional approaches, including focused English Language Development, Specially Designed Academic Instruction in English (SDAIE), scaffolding across content areas, structured English immersion, and determine communicative intent, particularly with students with low verbal abilities.	Cnt Les InM Asm Cmt	--[a]	AsD Fdk Cmt	●		Vid Cnt Les Cmt	Cnt	●		Les Vid Cmt	Les Cmt Cmp	●

(continued)

Table 2.1 (Continued)

TPE 1: Engaging and Supporting all Students in Learning	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
7. Provide students with opportunities to access the curriculum by incorporating the visual and performing arts, as appropriate to the content and context of learning.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	--[a]	--[a]
8. Monitor student learning and adjust instruction while teaching so that students continue to be actively engaged in learning.	Les InM Asm Cmt	Vid Cmt	AsD Fdk Cmt	●		Les Vid Cmt	Les Vid AsD Fdk Cmt	●		Vid	Les Asm AsD Cmt	●

Note. Cnt = Context; Les = Lesson/unit plan; InM = Instructional materials; Vid = Video; Asm = Assessments; Scr = Scoring criteria; AsD = Assessment data; Fdk = Feedback on student learning; Cmt = Commentary/narrative/reflection; Cmp = Classroom management plan.

● = weak evidence; ● = moderate evidence; ● = strong evidence.

Yellow highlights indicate a change to the linkage matrix from the Year 1 content validity workshop.

[a] The dashed lined indicates no evidence. The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [b]The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program.

Table 2.2. TPE 2: Evidence Types Mapped to TPA Components and Strength of Evidence Dashboard

TPE 2: Creating and Maintaining Effective Environments for Student Learning	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
1. Promote students' social-emotional growth, development, and individual responsibility using positive interventions and supports, restorative justice, and conflict resolution practices to foster a caring community where each student is treated fairly and respectfully by adults and peers.	--[a]	--[a]	--[a]	●[bc]		Cnt Les Vid Cmt	Vid Cmt	●		Vid	Cnt Cmp Cmt	●
2. Create learning environments (i.e., traditional, blended, and online) that promote productive student learning, encourage positive interactions among students, reflect diversity and multiple perspectives, and are culturally responsive.	--[a]	Vid Cmt	--[a]	●		Cnt Les Vid Cmt	--[a]	●		Vid Cmt	Cnt Cmp Cmt	●
3. Establish, maintain, and monitor inclusive learning environments that are physically, mentally, intellectually, and emotionally healthy and safe to enable all students to learn, and recognize and appropriately address instances of intolerance and harassment among students, such as bullying, racism, and sexism.	--[a]	Vid Cmt	--[a]	●[b]		Cnt Les Vid Cmt	Vid Cmt	●		Vid	Cmt Cmp	●
4. Know how to access resources to support students, including those who have experienced trauma, homelessness, foster care, incarceration, and/or are medically fragile.	--[a]	--[a]	--[a]	--[a]		Cnt Les Cmt	Cnt Les Cmt	●		--[a]	--[a]	--[a]
5. Maintain high expectations for learning with appropriate support for the full range of students in the classroom.	--[a]	Vid Cmt	--[a]	●		Les Vid Cmt	Les Vid Asm Fdk Cmt	●		Vid	Les Cmt Cmp	●
6. Establish and maintain clear expectations for positive classroom behavior and for student-to-student and student-to-teacher interactions by communicating classroom routines, procedures, and norms to students and families.	--[a]	Vid Cmt	--[a]	●		Vid Cmt Les	Vid Cmt	●		Vid	Cmt Cmp	●

Note. Cnt = Context; Les = Lesson/unit plan; InM = Instructional materials; Vid = Video; Asm = Assessments; Scr = Scoring criteria; AsD = Assessment data; Fdk = Feedback on student learning; Cmt = Commentary/narrative/reflection; Cmp = Classroom management plan.

● = weak evidence; ● = moderate evidence; ● = strong evidence.

Yellow highlights indicate a change to the linkage matrix from the Year 1 content validity workshop.

[a] The dashed lined indicates no evidence. The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [b] The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [c] No explicit requirement for the TPE by the TPA tasks/cycles; however, SMEs felt there was an implicit or indirect requirement for the TPE, and, thus, they assigned a sufficiency of evidence rating of '1' for weak evidence.

Table 2.3. TPE 3: Evidence Types Mapped to TPA Components and Strength of Evidence Dashboard

TPE 3: Understanding and Organizing Subject Matter for Student Learning	edTPA				CalTPA			FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength	Cycle 1	Cycle 2	Strength	SVP	TSP	Strength
1. Demonstrate knowledge of subject matter, including the adopted California State Standards and curriculum frameworks.	Les InM Asm Cmt	Vid Cmt	AsD Fdk Cmt	●	Les Vid Cmt	Les Vid Asm AsD Cmt	●	Les Vid Cmt	Les Asm Fbk Cmt	●
2. Use knowledge about students and learning goals to organize the curriculum to facilitate student understanding of subject matter and make accommodations and/or modifications as needed to promote student access to the curriculum.	Cnt Les InM Asm Cmt	--[a]	Fdk Cmt	●	Les Vid Cmt	Les Vid Asm AsD Cmt	●	Cnt Les Vid	Cnt Les Asm AsD Cmt	●
3. Plan, design, implement, and monitor instruction consistent with current subject-specific pedagogy in the content area(s) of instruction, and design and implement disciplinary and cross-disciplinary learning sequences, including integrating the visual and performing arts as applicable to the discipline.	Les InM Asm Cmt	Vid Cmt	Asm AsD Fdk Cmt	●[b]	Les Vid Cmt InM	Les Vid Asm AsD Cmt	●	Les Vid Cmt	Les Cmt	●[b]
4. Individually and through consultation and collaboration with other educators and members of the larger school community, plan for effective subject matter instruction and use multiple means of representing, expressing, and engaging students to demonstrate their knowledge.	--[a]	--[a]	--[a]	●[bc]	--[a]	--[a]	●[bc]	--[a]	Les Asm Cmt	●[b]
5. Adapt subject matter curriculum, organization, and planning to support the acquisition and use of academic language within learning activities to promote the subject matter knowledge of all students, including the full range of English learners, Standard English learners, students with disabilities, and students with other learning needs in the least restrictive environment.	Les InM Asm Cmt	--[a]	AsD Cmt	●	Les Vid Cmt InM	Les Vid Asm AsD Cmt InM	●	Les Vid	Cnt Les Asm Cmt	●
6. Use and adapt resources, standards-aligned instructional materials, and a range of technology, including assistive technology, to facilitate students' equitable access to the curriculum.	--[a]	--[a]	--[a]	●[bc]	Cmt	Les Cmt	●	Les	Les Cmt	●

(continued)

Table 2.3 (Continued)

TPE 3: Understanding and Organizing Subject Matter for Student Learning	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
7. Model and develop digital literacy by using technology to engage students and support their learning, and promote digital citizenship, including respecting copyright law, understanding fair use guidelines and the use of Creative Commons license, and maintaining Internet security.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	--[a]	--[a]
8. Demonstrate knowledge of effective teaching strategies aligned with the internationally recognized educational technology standards.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	●[bc]		--[a]	--[a]	●[bc]

Note. Cnt = Context; Les = Lesson/unit plan; InM = Instructional materials; Vid = Video; Asm = Assessments; Scr = Scoring criteria; AsD = Assessment data; Fdk = Feedback on student learning; Cmt = Commentary/narrative/reflection; Cmp = Classroom management plan.

● = weak evidence; ● = moderate evidence; ● = strong evidence.

Yellow highlights indicate a change to the linkage matrix from the Year 1 content validity workshop.

[a] The dashed lined indicates no evidence. The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [b] The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [c] No explicit requirement for the TPE by the TPA tasks/cycles; however, SMEs felt there was an implicit or indirect requirement for the TPE, and, thus, they assigned a sufficiency of evidence rating of '1' for weak evidence.

Table 2.4. TPE 4: Evidence Types Mapped to TPA Components and Strength of Evidence Dashboard

TPE 4: Planning Instruction and Designing Learning Experiences for All Students	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
1. Locate and apply information about students' current academic status, content- and standards-related learning needs and goals, assessment data, language proficiency status, and cultural background for both short-term and long-term instructional planning purposes.	Cnt Les InM Cmt	Vid Cmt	Cmt	●		Cnt Les Vid Cmt	Cnt Les Asm Cmt Vid	●		Cnt Les Vid	Cnt Les Cmt Asm	●
2. Understand and apply knowledge of the range and characteristics of typical and atypical child development from birth through adolescence to help inform instructional planning and learning experiences for all students.	Cmt	--[a]	--[a]	●		Cnt Vid Cmt	Cnt Les Asm Cmt	●		Cnt Les Vid Cmt	Cnt Les Cmt	●
3. Design and implement instruction and assessment that reflects the interconnectedness of academic content areas and related student skills development in literacy, mathematics, science, and other disciplines across the curriculum, as applicable to the subject area of instruction.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	Cnt Les Asm Cmt	●[b]
4. Plan, design, implement and monitor instruction, making effective use of instructional time to maximize learning opportunities and provide access to the curriculum for all students by removing barriers and providing access through instructional strategies that include: • appropriate use of instructional technology, including assistive technology; • applying principles of UDL and MTSS; • use of developmentally, linguistically, and culturally appropriate learning activities, instructional materials, and resources for all students, including the full range of English learners; • appropriate modifications for students with disabilities in the general education classroom; • opportunities for students to support each other in learning; and • use of community resources and services as applicable.	--[a]	Vid Cmt	Cmt	●[b]		Les Vid Cmt InM	Les Vid Asm Cmt	●		Vid	Les Asm AsD Cmt	●

(continued)

Table 2.4 (Continued)

TPE 4: Planning Instruction and Designing Learning Experiences for All Students	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
5. Promote student success by providing opportunities for students to understand and advocate for strategies that meet their individual learning needs and assist students with specific learning needs to successfully participate in transition plans (e.g., IEP, IFSP, ITP, and 504 plans.)	--[a]	--[a]	--[a]	●[bc]		--[a]	Les Vid Asm Scr Cmt	●		--[a]	--[a]	--[a]
6. Access resources for planning and instruction, including the expertise of community and school colleagues through in-person or virtual collaboration, co-teaching, coaching, and/or networking.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	--[a]	--[a]
7. Plan instruction that promotes a range of communication strategies and activity modes between teacher and student and among students that encourage student participation in learning.	--[a]	Vid Cmt	--[a]	●		Les Cmt	Les Vid Cmt	●		Les	Les Cmt	●
8. Use digital tools and learning technologies across learning environments as appropriate to create new content and provide personalized and integrated technology-rich lessons to engage students in learning, promote digital literacy, and offer students multiple means to demonstrate their learning.	--[a]	--[a]	--[a]	--[a]		--[a]	Les Cmt Vid InM	●		--[a]	Les Cmt InM	●[b]

Note. Cnt = Context; Les = Lesson/unit plan; InM = Instructional materials; Vid = Video; Asm = Assessments; Scr = Scoring criteria; AsD = Assessment data; Fdk = Feedback on student learning; Cmt = Commentary/narrative/reflection; Cmp = Classroom management plan.

● = weak evidence; ● = moderate evidence; ● = strong evidence.

Yellow highlights indicate a change to the linkage matrix from the Year 1 content validity workshop.

[a] The dashed lined indicates no evidence. The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [b] The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [c] No explicit requirement for the TPE by the TPA tasks/cycles; however, SMEs felt there was an implicit or indirect requirement for the TPE, and, thus, they assigned a sufficiency of evidence rating of '1' for weak evidence.

Table 2.5. TPE 5: Evidence Types Mapped to TPA Components and Strength of Evidence Dashboard

TPE 5: Assessing Student Learning	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
1. Apply knowledge of the purposes, characteristics, and appropriate uses of different types of assessments (e.g., diagnostic, informal, formal, progress-monitoring, formative, summative, and performance) to design and administer classroom assessments, including use of scoring rubrics.	--[a]	Vid Cmt	Asm Scr AsD Fdk Cmt	●		Les Vid Cmt	Les Vid Asm Scr AsD Cmt	●		Les	Les Asm AsD Cmt	●
2. Collect and analyze assessment data from multiple measures and sources to plan and modify instruction and document students' learning over time.	--[a]	--[a]	Cmt	●		Cmt	Vid Asm AsD Cmt	●		--[a]	Asm AsD Cmt	●
3. Involve all students in self-assessment and reflection on their learning goals and progress and provide students with opportunities to revise or reframe their work based on assessment feedback.	--[a]	--[a]	--[a]	●[bc]		--[a]	Vid Scr AsD Cmt	●		--[a]	--[a]	--[a]
4. Use technology as appropriate to support assessment administration, conduct data analysis, and communicate learning outcomes to students and families.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	●[bc]		--[a]	--[a]	●[bc]
5. Use assessment information in a timely manner to assist students and families in understanding student progress in meeting learning goals.	--[a]	--[a]	Fdk Cmt	●[b]		--[a]	Vid AsD Fdk Cmt	●[b]		--[a]	AsD Fdk Cmt	●
6. Work with specialists to interpret assessment results from formative and summative assessments to distinguish between students whose first language is English, English learners, Standard English learners, and students with language or other disabilities.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	--[a]	--[a]

(continued)

Table 2.5 (Continued)

TPE 5: Assessing Student Learning	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
7. Interpret English learners' assessment data to identify their level of academic proficiency in English as well as in their primary language, as applicable, and use this information in planning instruction.	--[a]	--[a]	--[a]	--[a]		--[a]	Cnt Les Cmt	●[b]		Cnt Les Cmt	Cnt Les Cmt	●[b]
8. Use assessment data, including information from students' IEP, IFSP, ITP, and 504 plans, to establish learning goals and to plan, differentiate, make accommodations and/or modify instruction.	--[a]	--[a]	AsD Fdk Cmt	●		Cnt Les Vid Cmt	Cnt Les Vid AsD Cmt	●		Cnt Les Cmt	Cnt Les Asm Cmt	●

Note. Cnt = Context; Les = Lesson/unit plan; InM = Instructional materials; Vid = Video; Asm = Assessments; Scr = Scoring criteria; AsD = Assessment data; Fdk = Feedback on student learning; Cmt = Commentary/narrative/reflection; Cmp = Classroom management plan.

● = weak evidence; ● = moderate evidence; ● = strong evidence.

Yellow highlights indicate a change to the linkage matrix from the Year 1 content validity workshop.

[a] The dashed lined indicates no evidence. The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [b] The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [c] No explicit requirement for the TPE by the TPA tasks/cycles; however, SMEs felt there was an implicit or indirect requirement for the TPE, and, thus, they assigned a sufficiency of evidence rating of '1' for weak evidence.

Table 2.6. TPE 6: Evidence Types Mapped to TPA Components and Strength of Evidence Dashboard

TPE 6: Developing as a Professional Educator	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
1. Reflect on their own teaching practice and level of subject matter and pedagogical knowledge to plan and implement instruction that can improve student learning.	--[a]	Cmt	Cmt	●		Vid Cmt	Cmt	●		Cmt	Cmt	●
2. Recognize their own values and implicit and explicit biases, the ways in which these values and implicit and explicit biases may positively and negatively affect teaching and learning, and work to mitigate any negative impact on the teaching and learning of students. They exhibit positive dispositions of caring, support, acceptance, and fairness toward all students and families, as well as toward their colleagues.	--[a]	Vid Cmt	--[a]	●[b]		--[a]	Vid Cmt	●[b]		Vid	Cmp Cmt	●[b]
3. Establish professional learning goals and make progress to improve their practice by routinely engaging in communication and inquiry with colleagues.	--[a]	--[a]	--[a]	--[a]		--[a]	Cmt	●[b]		--[a]	Cmt	●[b]
4. Demonstrate how and when to involve other adults and to communicate effectively with peers and colleagues, families, and members of the larger school community to support teacher and student learning.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	--[a]	●[bc]
5. Demonstrate professional responsibility for all aspects of student learning and classroom management, including responsibility for the learning outcomes of all students, along with appropriate concerns and policies regarding the privacy, health, and safety of students and families. Beginning teachers conduct themselves with integrity and model ethical conduct for themselves and others.	--[a]	--[a]	--[a]	--[a]		Vid Cmt	--[a]	●[b]		Cmt	Cnt Cmt Cmp	●[b]

(continued)

Table 2.6 (Continued)

TPE 6: Developing as a Professional Educator	edTPA					CalTPA				FAST		
Beginning Teachers:	Task 1	Task 2	Task 3	Strength		Cycle 1	Cycle 2	Strength		SVP	TSP	Strength
6. Understand and enact professional roles and responsibilities as mandated reporters and comply with all laws concerning professional responsibilities, professional conduct, and moral fitness, including the responsible use of social media and other digital platforms and tools.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	--[a]	--[a]
7. Critically analyze how the context, structure, and history of public education in California affects and influences state, district, and school governance as well as state and local education finance.	--[a]	--[a]	--[a]	--[a]		--[a]	--[a]	--[a]		--[a]	--[a]	--[a]

Note. Cnt = Context; Les = Lesson/unit plan; InM = Instructional materials; Vid = Video; Asm = Assessments; Scr = Scoring criteria; AsD = Assessment data; Fdk = Feedback on student learning; Cmt = Commentary/narrative/reflection; Cmp = Classroom management plan.

• = weak evidence; • = moderate evidence; • = strong evidence.

Yellow highlights indicate a change to the linkage matrix from the Year 1 content validity workshop.

[a] The dashed lined indicates no evidence. The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program. [b]The Commission requires that all TPEs must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program.

Comparisons across TPAs

A summary of findings by TPE domain are provided below. It should be noted that the Commission requires that all of the teaching performance expectations must be measured by accredited teacher preparation programs and any TPEs not specifically measured by the TPA must be assessed elsewhere in the program.

TPE 1: Engaging and Supporting all Students in Learning

- All models had at least one TPE element receive a "strong evidence" rating
- TPE element 1.8 ("*Monitor student learning and adjust instruction while teaching so that students continue to be actively engaged in learning"*) received a "strong evidence" rating for all three models
- CalTPA had the most TPE elements (3 of 8) receive a "strong evidence" rating
- None of the models had more than two of the eight elements receive "no" or "weak" evidence ratings

TPE 2: Creating and Maintaining Effective Environments for Student Learning

- FAST had the most TPE elements (2 of 6) receive a "strong evidence" rating
- edTPA was the only model that did not receive a "strong evidence" rating on at least one TPE element, although it did meet the requirements of the ADS by addressing "key aspects" of multiple TPE elements within this domain

TPE 3: Understanding and Organizing Subject Matter for Student Learning

- All models had the same three TPE elements receive a "strong evidence" rating (i.e., TPEs 3.1, 3.2, and 3.5)

TPE 4: Planning Instruction and Designing Learning Experiences for All Students

- FAST had the most TPE elements (2 of 8) receive a "strong evidence" rating
- edTPA was the only model that did not receive a "strong evidence" rating on at least one TPE element, although it did meet the requirements of the ADS by addressing "key aspects" of multiple TPE elements within this domain

TPE 5: Assessing Student Learning

- TPE element 5.1 ("Apply knowledge of the purposes, characteristics, and appropriate uses of different types of assessments (e.g., diagnostic, informal, formal, progress-monitoring, formative, summative, and performance) to design and administer classroom assessments, including use of scoring rubrics") received a "strong evidence" rating for all three models
- Both CalTPA and FAST had half of the elements receive a rating of "moderate" or "strong" evidence
- CalTPA had the fewest elements receive a rating of "no" evidence (1 of 8)

TPE 6: Developing as a Professional Educator

- All models received a "strong" evidence rating on one element in this TPE domain (TPE 6.1: ("Reflect on their own teaching practice and level of subject matter and pedagogical knowledge to plan and implement instruction that can improve student learning")
- For all three models, all elements in this domain, except TPE 6.1, received a rating of "no" or "weak" evidence

Discussion

The purpose of Activity 2 was to further investigate Claim 1. Activity 2 delves deeper into the part of Claim 1 that states, "*The TPA models are sufficiently comparable in their assessment and weighting of the Commission-adopted TPEs."* Activity 2 took the format of a content validity workshop with panels of teacher preparation experts (i.e., SMEs). Each panel mapped the components of each TPA (i.e., tasks or cycles) to each TPE element by identifying the types of evidence candidates are required to submit to demonstrate the KSAs specified in the TPE element. Then, the SMEs evaluated how thoroughly the collective set of evidence required by each model assesses the KSAs in each TPE element. These linkages were updated in summer 2019 given changes that occurred to the models following the spring 2018 content validity workshop.

ADS 1(a) requires all TPA tasks/cycles to measure "*two or more*" TPE domains. All models exceed this requirement by having all components (i.e., tasks or cycles) of each model assess *three or more* of the six TPE domains. In addition, all models were judged by the SMEs to require strong evidence of the KSAs specified by one or more TPE elements for at least four of the six TPE domains. Moreover, across all models, TPE 3 (Understanding and Organizing Subject Matter for Student Learning) received the strongest linkages to evidence requirements and TPE 6 (Developing as a Professional Educator) received the fewest linkages to evidence requirements. This indicates that all three models tend to do the best job of assessing the breadth and depth of TPE 3 and the poorest job of assessing the breadth and depth of TPE 6; the SMEs commented that the KSAs described in TPE 6 are difficult to measure via a performance assessment.

Overall, CalTPA and FAST were similar with regard to the frequency with which evidence requirements were linked to TPEs, although there tended to be slightly more evidence linkages for CalTPA (particularly for TPE 2: Creating and Maintaining Effective Environments for Student Learning and TPE 4: Planning Instruction and Designing Learning Experiences for All Students). However, the strength of evidence ratings were, overall, very similar for both CalTPA and FAST. The edTPA tended to have slightly fewer evidence linkages and slightly lower strength of evidence ratings, particularly for TPE domains 2 and 4. This may be due in part to edTPA being a national performance assessment; whereas CalTPA and FAST were developed specifically to address the California TPEs. Thus, CalTPA and FAST are more attuned to the specific terminology used in California to describe the KSAs of beginning teachers, whereas the terminology used in edTPA is less California-centric.

Conclusion

This activity provides an independent, empirical investigation on the content validity of each of the models. The findings demonstrate that each model adheres to *Assessment Design Standard* 1(a) by demonstrating that each task/cycle for each model substantively assesses two or more TPE domains. The models are also comparable in that TPE 3 (Understanding and Organizing Subject Matter for Student Learning) was the TPE domain assessed most thoroughly by all three models and TPE 6 (Developing as a Professional Educator) was the TPE domain assessed least thoroughly by all three models (i.e., TPE domain 6 had the fewest and weakest linkages to assessment components). Thus, it is important for the programs to ensure that TPE 6 is being addressed by the programs through other means. The models are also comparable in that commentary/narrative/reflection is the evidence type most frequently required by the models to demonstrate candidates' KSAs.

The results from this activity indicate that CalTPA and FAST tend to be more comparable with one another than either is with edTPA, and that edTPA does not have any components (tasks) for which there was strong evidence for the assessment of TPE 2 (Creating and Maintaining Effective Environments for Student Learning) and TPE 4 (Planning Instruction and Designing Learning Experiences for All Students), as judged by the SMEs. However, the SMEs agreed that there was moderate evidence for the assessment of TPEs 2 and 4 by edTPA, meaning that edTPA assessed key aspects of the TPE elements in these domains, but not the full breadth or depth of these TPEs.

In conclusion, there are some differences in the emphasis and measurement of TPEs across the TPA models; however, there is more comparability than dissimilarity across models, particularly between FAST and CalTPA. This provides partial but not full support for the claim that the TPA models are sufficiently comparable in their assessment and weighting of the TPEs (Claim 1). The *Assessment Design Standards* state that each TPA task must be related to two or more major domains of the TPEs and that the tasks and rubrics must collectively address key aspects of the six TPE domains (ADS 1a). However, the *Assessment Design Standards* do not specify which of the 45 TPE elements or how many of them each model must address. Thus, it is perhaps not surprising that a mapping of TPE elements to TPA components revealed some differences across the three TPA models.

Chapter 3: Comparison of Stakeholder Input across TPA Models (Activity 3)

Randy Knebel & Andrea L. Sinclair

Introduction

The primary purpose of Activity 3 was to investigate Claim 2, which is, *"The guidance and supports (e.g., guide/manual/handbook and other resources) provided by model sponsors to candidates and teacher preparation faculty are sufficiently clear and detailed to ensure that the model is implemented as designed and intended."* The findings from this activity help to inform the extent to which this claim is substantiated for each of the three TPA models. To investigate this claim we gathered stakeholder perceptions through on-line surveys. Candidates and program coordinators are two of the most important stakeholder groups for this purpose.[19] Consequently, surveys were administered to gather perceptions of the usefulness of the guidance and supports provided to (a) candidates and (b) program coordinators. If candidates and program coordinators find the guidance and supports provided by the model sponsors clear, detailed, and useful, then that should lay the groundwork for models to be implemented as designed and intended.

This activity also helps to further inform Claim 1, which is, "*The TPA models are sufficiently comparable in their representation of the Commission's Assessment Design Standards (ADS) and in their assessment and weighting of the Commission-adopted Teaching Performance Expectations (TPEs)*." Claim 1 essentially pertains to the content validity of the TPA models. To assess this claim, these same surveys included items regarding candidates' and program coordinators' perceptions of the validity of the TPA models—that is, the face validity of the TPA models. In this sense, the survey findings provide information on the validity of the TPA models from the perspective of candidates and program coordinators.

Method

Context/Background. A Candidate Survey was administered to candidates who recently completed their submissions. A Coordinator Survey was simultaneously administered to program coordinators at the candidate institutions. These surveys were also administered in Year 1 (2017-2018) of the comparability study and those findings are presented in the Year 1 report (Sinclair & Thacker, 2018). In Year 1, the CalTPA model sponsors developed their own Candidate and Coordinator Surveys. HumRRO reviewed those CalTPA-developed surveys and found that the CalTPA surveys contained the same types of items that HumRRO planned to include on its Candidate and Coordinator Surveys. Thus, to minimize survey burden, rather than develop a separate HumRRO-developed Candidate Survey and Coordinator Survey, HumRRO coordinated with the CalTPA representatives to include some additional items on the CalTPA-developed surveys and to obtain data from those CalTPA-administered surveys to inform Activity 3. HumRRO developed parallel surveys for edTPA and FAST. The surveys included items that appeared on surveys for all three models *and* some items that were similar (e.g., usefulness of specific supports and resources), but were contextualized for each model to use TPA-specific terminology. Because of this, direct item-to-item comparison of results across models was not always possible.

[19] The term "program coordinators" is not used at Fresno State for the FAST model. The counterparts at Fresno State are referred to as "university coaches." For the sake of simplicity, the term "program coordinators" is being used in this report to refer to those responsible for preparing candidates for all three models.

In Year 2, HumRRO learned that CalTPA changed their surveys somewhat from Year 1 to Year 2. As a result, the Year 2 survey results presented in this chapter reflect some item differences between the CalTPA surveys and the edTPA and FAST surveys, which remained unchanged from Year 1 to Year 2. Thus, the item level results are not as comparable as we might like. Nonetheless, the survey items across all three models target the same fundamental topics—that is, (a) clarity/usefulness of supports and materials and (b) perceived validity of the TPA. And, thus, comparisons at the topic level are warranted.

Survey Format. The online surveys comprised Likert-scale items. The survey items can be found in the item-level tables of results in Appendices 3.A (FAST Candidate Survey), 3.B (edTPA Candidate Survey), 3.C (CalTPA Candidate Survey), and 3.D (which includes FAST, edTPA, and CalTPA Coordinator Surveys).[20]

Administration. For FAST and edTPA, HumRRO provided the platform for survey administration. For FAST, the survey URL was emailed directly to candidates and coordinators by the model sponsor. For edTPA, the survey URL was emailed to program coordinators by the model sponsor; however, for candidates the survey URL was emailed to coordinators, rather than directly to candidates, and the coordinators were asked to forward the survey URL to the candidates in their program. This is the standard procedure that the Stanford Center for Assessment, Learning and Equity (SCALE) uses for communicating with candidates and to which we adhered.[21] The surveys for edTPA and FAST were launched on/about 17 April 2019 and closed on 15 May 2019. The Evaluation Systems group of Pearson (ES) administered the CalTPA surveys. ES launched the Coordinator Survey on 15 March 2019 and the Candidate Survey on 21 March 2019, and HumRRO received the data extract on 16 May 2019.[22] The data for CalTPA was provided to HumRRO by ES.

Results

Frequency distributions of participant responses are summarized in tables and figures below. Candidate responses are compared within and across models (where possible). Survey responses are depicted graphically (i.e., figures of color-coded bar charts) to illustrate similarities and differences between the responses of stakeholders from each model. The item-level frequency distribution tables on which the figures are based are located in Appendices 3.A – 3.D.

Response Rates and Data Cleaning

Response rates are shown in Table 3.1 and Table 3.2 for candidates and coordinators, respectively. Upon receiving the data, standard data cleaning checks were conducted (i.e., missing data, out-of-range values, and irregular patterns in the data). The checks indicated sensible and expected patterns in the data. Cases were removed for which no responses were received (see Tables 3.1 and 3.2). Analyses were conducted in SPSS 25.0. Valid percentages of responses (i.e., excluding missing responses) are reported in all results. Candidate response rates were higher in Year 2 than in Year 1 for FAST and edTPA. However, the overall response rate was still quite low for edTPA at 7 percent; this is likely due, at least in part, to the fact that the survey URL was not sent directly to candidates, but to coordinators who were then asked to

[20] Appendices for this report are in Volume II: Appendices.
[21] SCALE designed and supports the edTPA.
[22] Although HumRRO received the data extract on 16 May 2019, we learned later that the CalTPA candidate and Coordinator Surveys remained “live” for a short period of time after this date. To confirm that this did not result in a truncated data set, we verified with the model sponsor that our data extract for the Candidate Survey was based on a near complete respondent sample (i.e., just seven fewer cases in our analysis sample than CalTPA’s analysis sample) and a complete dataset for the Coordinator Survey.

forward the survey URL to their candidates. The candidate response rate was notably lower in Year 2 than in Year 1 for CalTPA (i.e., 27.3% in Year 1 compared to 10.4% in Year 2). Response rates were considerably higher for coordinators than for candidates. The highest coordinator response rate was for FAST at 77 percent and lowest for edTPA at 40 percent.

Table 3.1. Survey Response Rate for Candidates

	FAST		edTPA		CalTPA	
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Total N	~130[ab]	236[b]	~4,175[c]	~4,743[c]	1,736	3,917
N Responded	42	106	103	363	505	407
N Valid Cases	40	94	96	330	474	407
N (%) Missing	2 (4.8)	12 (11.3)	7 (6.8)	33 (9.1)	31 (6.1)	0 (0)
Response Rate (%)	~30.8	44.9	~2.3[d]	~7.0[d]	27.3	10.4

[a]Approximate number reported by model sponsor.
[b]Survey sent to teacher candidates in final student teaching who completed both FAST tasks in the academic year.
[c]Number comes from the California Biannual Summary Report and represents all submissions from all examinees; it is important to note that this number includes retakes and thus is an overrepresentation of the actual number of candidates.
[d]It is important to note that the edTPA Candidate Survey was not directly delivered to candidates. Per SCALE's standard procedure, the URL link for the Candidate Survey was emailed to edTPA Program Coordinators who were instructed to forward the survey link to their candidates.

Table 3.2. Survey Response Rates for Coordinators

	FAST		edTPA		CalTPA	
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Total N	39	~70[a]	75[b]	58[b]	36	93
N Responded	25	65	31	26	22	54
N Valid Cases	23	54	24	23	20	50
N (%) Missing	2 (8.0)	11 (16.9)	7 (22.6)	3 (11.5)	2 (9.1)	4 (7.4)
Response Rate (%)	59.0	~77.1	32.0	39.7	55.6	53.8

[a]Approximate number reported by model sponsor.
[b]Not including coordinators for whom the email with the survey URL was bounced back as undeliverable.

Respondent Characteristics

Candidates. As shown in Table 3.3, the majority of respondents for both edTPA and FAST were female (68.2% and 73.5%, respectively). Slightly less than half of candidate respondents for edTPA and FAST reported that they were White (non-Hispanic), as shown in Table 3.4. The majority of FAST and CalTPA candidates reported seeking a multiple subject credential; however, the majority of edTPA candidates reported that they were seeking a single subject credential, as shown in Table 3.5.

Table 3.3. Distribution of Gender on Candidate Survey

Gender	FAST (%)	edTPA (%)	CalTPA (%)
Male	26.5	28.0	NA[a]
Female	73.5	68.2	NA
Non-Binary	0.0	1.3	NA
Other	0.0	2.5	NA

[a] NA = Not Available. Gender was not collected on the CalTPA survey.

Table 3.4. Distribution of Candidate Race/Ethnicity on Candidate Survey

Race/Ethnicity[a]	FAST (%)	edTPA (%)	CalTPA (%)
African American/Black	1.1	1.2	NA [b]
Filipino American/Filipino	1.1	3.0	NA
Latino/Latin American/Puerto Rican/Other Hispanic	9.6	6.7	NA
Mexican American/Chicano	22.3	11.5	NA
White (non-Hispanic)	44.7	40.3	NA
Other	3.2	3.6	NA
Choose not to respond	7.5	11.2	NA

[a] Infrequently selected races/ethnicities are not included in the table.
[b] NA = Not Available. Race/ethnicity was not collected on the CalTPA survey.

Table 3.5. Candidate Distribution of Type of Preliminary Teaching Credential

Teaching Credential	FAST (%)	edTPA (%)	CalTPA (%)
Multiple Subject	51.6	40.0	60.0
Single Subject	45.2	58.5	40.0
Education Specialist Credential	2.2	NA	NA
Other	1.1	1.5	NA

Coordinators. Although gender information was not collected on the CalTPA Coordinator Survey, FAST and edTPA coordinators most frequently reported they were female (79.6% and 70.0%, respectively), as shown in Table 3.6. Aside from those respondents who chose not to indicate their race/ethnicity (roughly 1 out of every 5 respondents), the majority of coordinators for FAST and edTPA reported they were White (53.7% and 47.8%, respectively; see Table 3.7). Race/ethnicity was not collected for coordinators on the CalTPA survey. Finally, as shown in Table 3.8, well over half of the FAST respondents were in their position for 3 years or less (69.4%). In comparison, 43 percent of edTPA coordinators and 44 percent of CalTPA coordinators have been in their position for this length of time.

Table 3.6. Distribution of Gender on Coordinator Survey

Gender	FAST (%)	edTPA (%)	CalTPA (%)
Male	20.4	20.0	NA[a]
Female	79.6	70.0	NA
Non-Binary	0.0	0.0	NA
Other	0.0	10.0	NA

[a] NA = Not Available. Gender was not collected on the CalTPA survey.

Table 3.7. Distribution of Coordinator Race/Ethnicity

Race/Ethnicity[a]	FAST (%)	edTPA (%)	CalTPA (%)
African American/Black	0.0	0.0	NA[b]
Filipino American/Filipino	0.0	4.4	NA
Latino/Latin American/Puerto Rican/Other Hispanic	3.7	8.7	NA
Mexican American/Chicano	14.8	4.4	NA
Native American/American Indian/Alaskan Native	1.9	4.4	NA
White (non-Hispanic)	53.7	47.8	NA
Other	1.9	4.4	NA
Choose not to respond	20.4	17.4	NA

[a] Infrequently selected races/ethnicities are not included in the table.
[b] NA = Not Available. Race/ethnicity was not collected on the CalTPA survey.

Table 3.8. Frequency Distribution of Coordinators' Length of Time in Present Position

Length of Time	FAST (%)	edTPA (%)	CalTPA (%)
One year or less	44.9	21.7	12.5
2-3 years	24.5	21.7	31.2
4-5 years	4.1	17.3	16.6
6-10 years	18.4	30.3	8.4
More than 10 years	8.2	8.6	23.1

Candidate Survey Results

The data below are presented by model and broken down by TPA component, where applicable.[23] The three models differ in structure and terminology, including component names. For example, the FAST Tasks are referred to as the Site Visitation Project (SVP) and the Teaching Sample Project (TSP). The edTPA has three Tasks: Planning (Task 1), Instruction (Task 2), and Assessment (Task 3); it should be noted that for the multiple-subject credential there is an assessment task for literacy and for mathematics, thereby resulting in two assessment tasks for the multiple-subject credential. Because the TPA models are structured differently, the data for all models could not be easily presented within the same figures. Therefore, each model is presented in a separate figure by TPA component, where applicable. Each of the models also used different terminology for its support materials and provided different types of supports. The surveys administered for FAST, edTPA, and CalTPA were all contextualized to use model-specific terminology. The wording of the survey questions is summarized in the figures due to space constraints. The full wording of the questions as they appear on the surveys can be found in the tables in Appendices 3.A through 3.D.

FAST Survey Results for Claim 2 (Clarity and Usefulness of Guidance/Supports)

The results presented in Figures 3.1 and 3.2 address Claim 2, which pertains to the clarity and ease of use of the support materials and guidance provided to the FAST candidates.

- 85.4 percent of FAST candidate respondents agreed or strongly agreed that, overall, the directions for the SVP were easy to understand. While still the majority, only 72.8 percent of FAST candidate respondents agreed or strongly agreed that, overall, the directions for the TSP were easy to understand.
- 92.7 percent of FAST candidate respondents reported that they agreed or strongly agreed that they understood what was expected in the reflection for the SVP. While still the majority, only 80.2 percent of respondents reported the same level of agreement for the TSP.
- 89.0 percent and 86.4 percent of respondents understood what they were asked to submit as evidence for the SVP and TSP, respectively.
- Similar percentages of respondents agreed or strongly agreed that the four levels of performance were clearly stated in the rubrics in the SVP (91.5%) and TSP (88.9%).
- Respondents similarly agreed or strongly agreed that the rubrics helped them prepare their submissions for the SVP (89.0%) and the TSP (90.1%).

[23] In 2018, the CalTPA Candidate Survey, on which the format of the FAST and edTPA Candidate Surveys were based, included candidate questions about *each* component (i.e., Cycle) of CalTPA. However, in 2019, the CalTPA Candidate Survey simply asked candidates to respond to questions about CalTPA overall, not separately about Cycle 1 and Cycle 2.

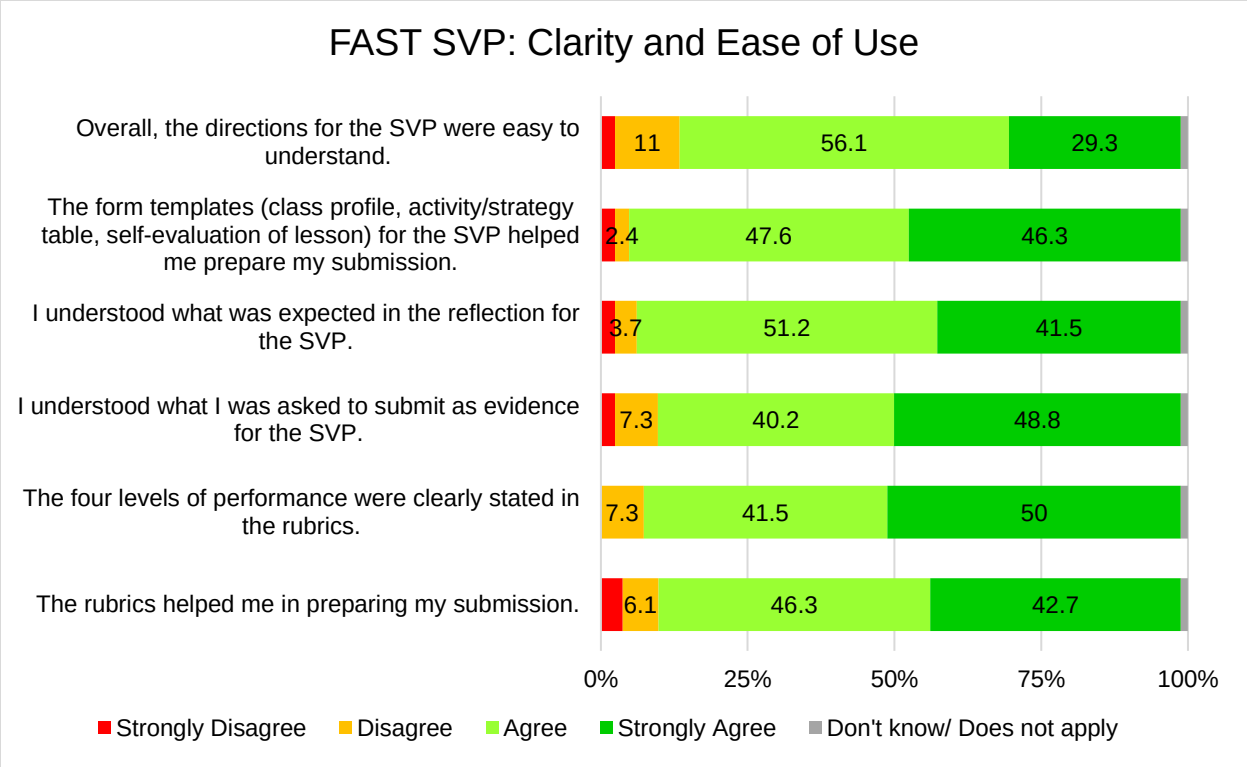

Figure 3.1. FAST SVP: Candidate perceptions of clarity and ease of use.

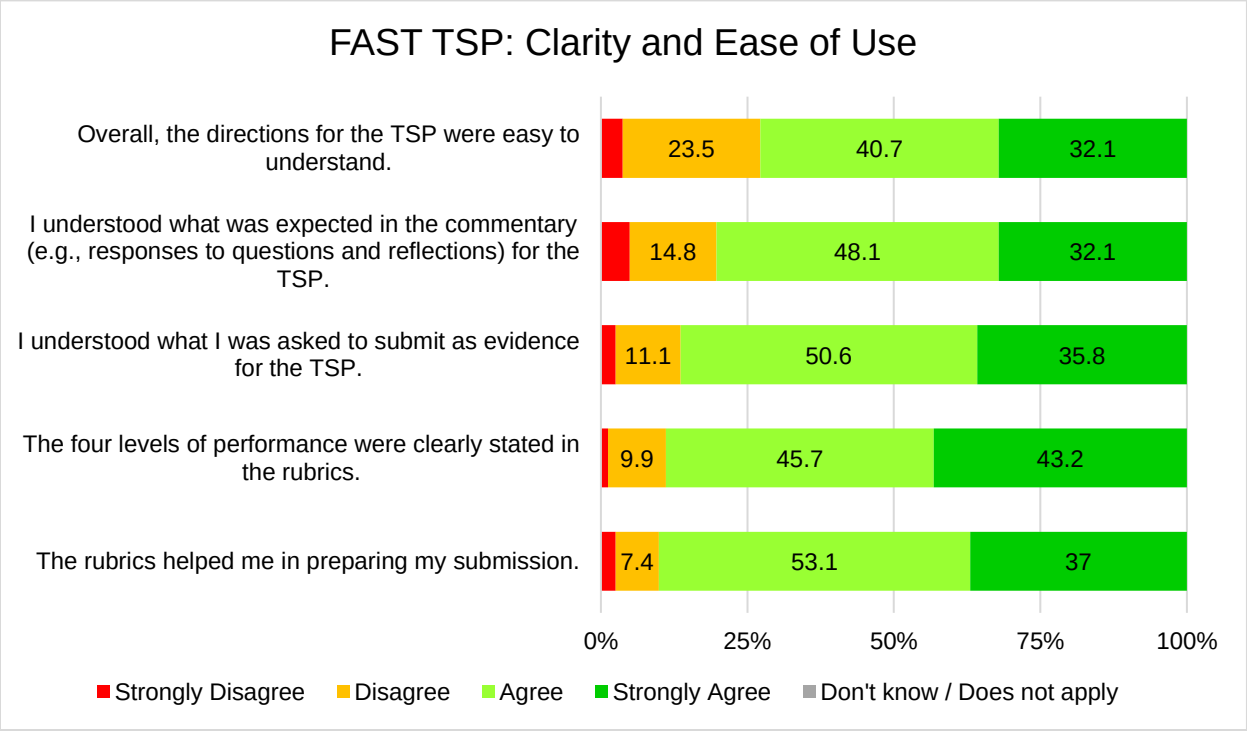

Figure 3.2. FAST TSP: Candidate perceptions of clarity and ease of use.

In another question, FAST candidates were asked to rate their level of satisfaction with the online system used to upload their submissions. The candidates did not find the TK20 online submission system very user friendly, with almost half of respondents saying that it was not a helpful resource (see Figure 3.3).

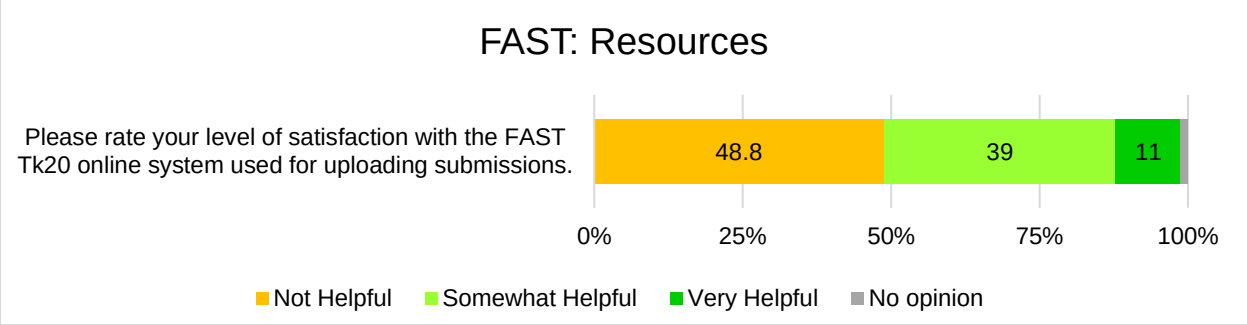

Figure 3.3. Candidate perceptions of FAST resources.

The majority of FAST candidate respondents agreed or strongly agreed that the FAST Manual provided sufficient information to assist them throughout the assessment process (see Figure 3.4).

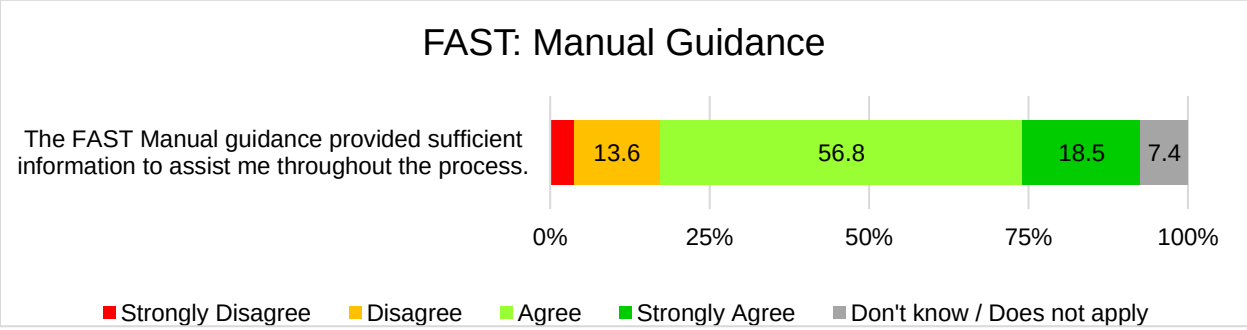

Figure 3.4. Candidate perceptions of FAST Manual guidance.

FAST Survey Results for Claim 1 (Perceived Validity)

The results presented below in Figures 3.5 and 3.6 address Claim 1; Claim 1 pertains to the perceived validity of FAST (i.e., candidates' perceptions that the knowledge, skills, and abilities—i.e., KSAs—assessed by FAST are emphasized in their program and that FAST provides an opportunity to demonstrate those KSAs).

- The majority of candidate respondents reported that the SVP and TSP provided them with sufficient opportunity to demonstrate their instructional knowledge, skills, and abilities (85.3% and88.5%, respectively).
- 88.9 percent of candidate respondents reported that the teaching knowledge, skills, and abilities assessed in the SVP are emphasized in their preparation program. This percentage is similar to the percentage of candidate respondents who reported that the teaching knowledge, skills, and abilities assessed in the TSP are emphasized in their preparation program (85.4%).

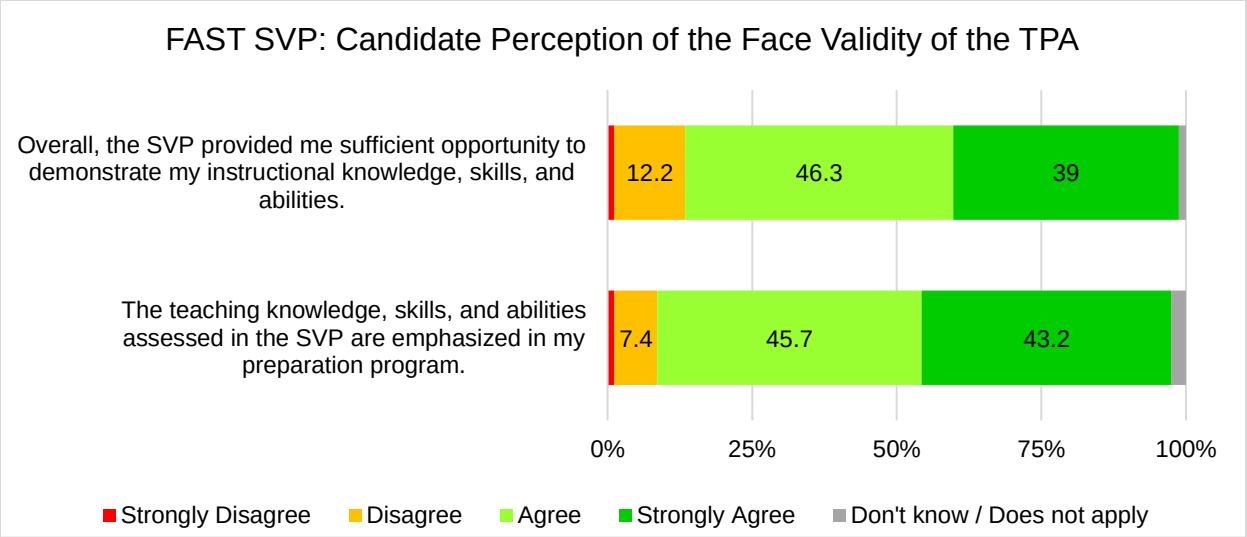

Figure 3.5. FAST SVP: Candidate perceptions of validity.

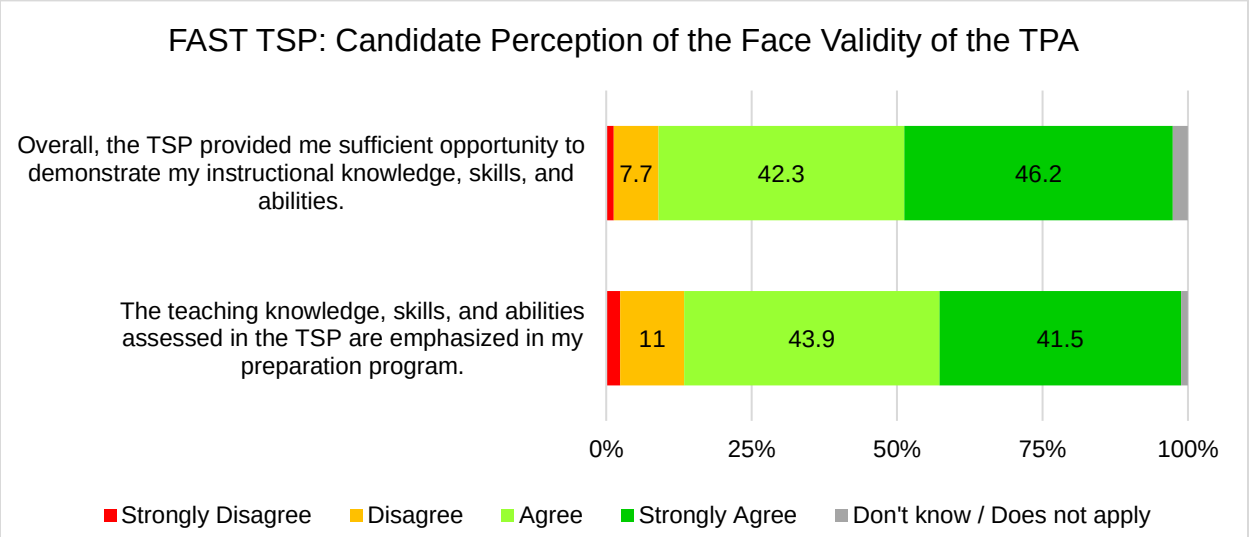

Figure 3.6. FAST TSP: Candidate perceptions of validity.

edTPA Survey Results for Claim 2 (Clarity and Usefulness of Guidance/Supports)

Figures 3.7, 3.8, and 3.9 display results from candidate respondents that address Claim 2, which pertains to the clarity and ease of use of the support materials and guidance provided to the candidates by the model sponsor. The following bullet points report comparisons among the edTPA components—i.e., Planning (Task 1), Instruction (Task 2), and Assessment (Task 3).

- While the majority of respondents reported that they agreed the directions for Tasks 1, 2, and 3 were easy to understand, over 35 percent of respondents strongly disagreed or disagreed that the directions for the Tasks were easy to understand. Notably, over

40 percent of edTPA candidate respondents strongly disagreed or disagreed that the directions for the Assessment Task (Task 3),[24] specifically, were easy to understand.

- For each Task, the majority of respondents understood what was expected in the commentary; however, roughly one out of every four candidates disagreed that they understood what was expected in the commentary.
- Three out of four respondents understood what they were asked to submit as evidence for each of the Tasks; however, roughly one out of every four candidates disagreed or strongly disagreed that they understood what they were asked to submit as evidence.
- While the majority of respondents indicated that the five levels of performance were clearly stated in the rubrics, nontrivial percentages of respondents (>25%) felt dissimilarly by indicating that they strongly disagreed or disagreed that the five levels of performance were clearly stated in the rubrics.
- Most candidates reported that the rubrics helped in preparing their submissions for each of the Tasks (> 70%); however, roughly one out of every four candidates reported that the rubrics did not help them in preparing their submissions.

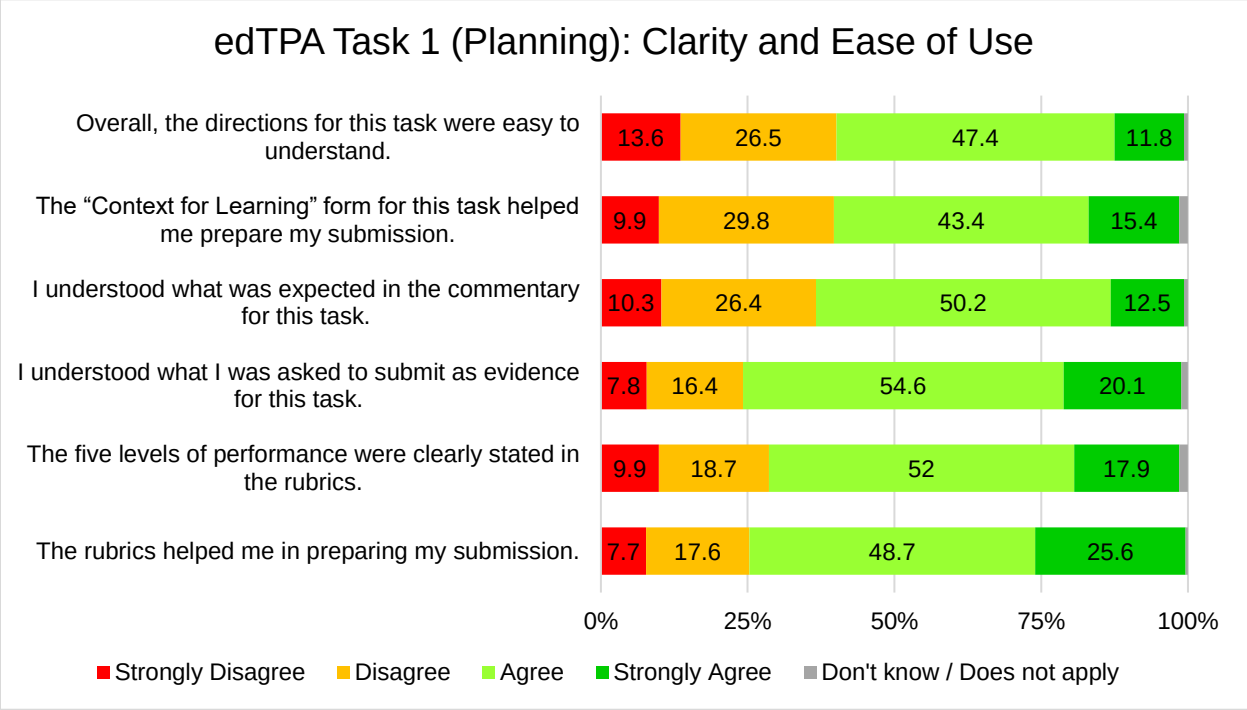

Figure 3.7. edTPA Task 1 (Planning): Candidate perceptions of clarity and ease of use.

[24] And Task 4 for the multiple-subject credential.

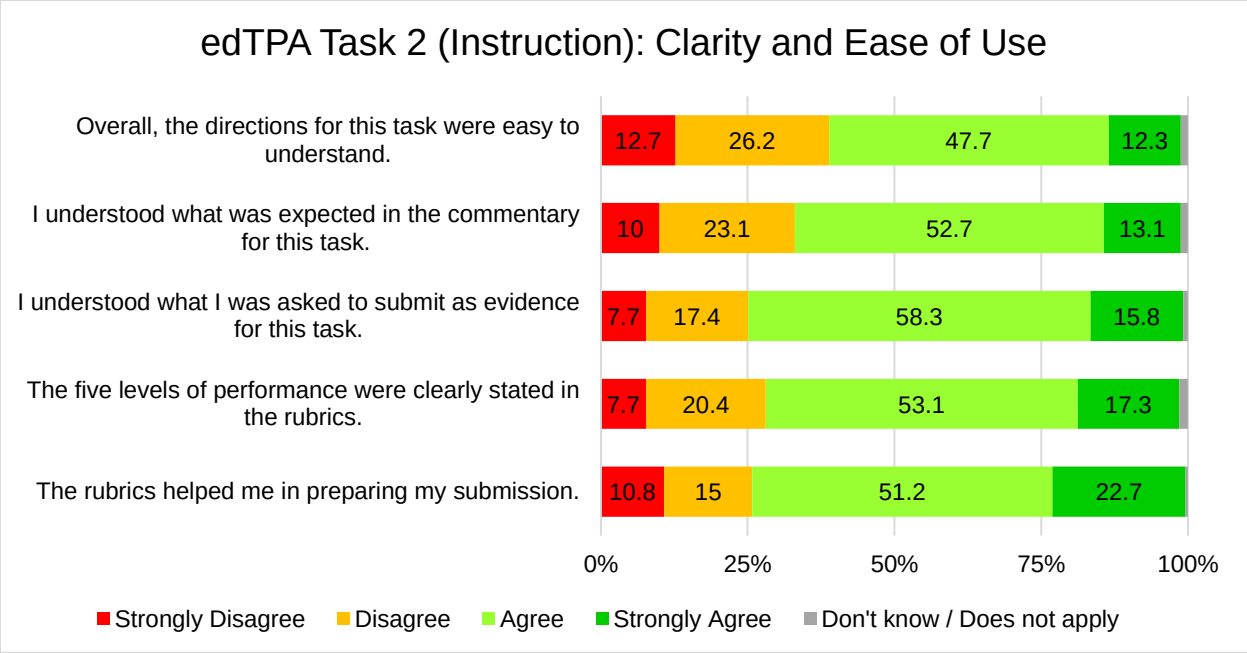

Figure 3.8. edTPA Task 2 (Instruction): Candidate perceptions of clarity and ease of use.

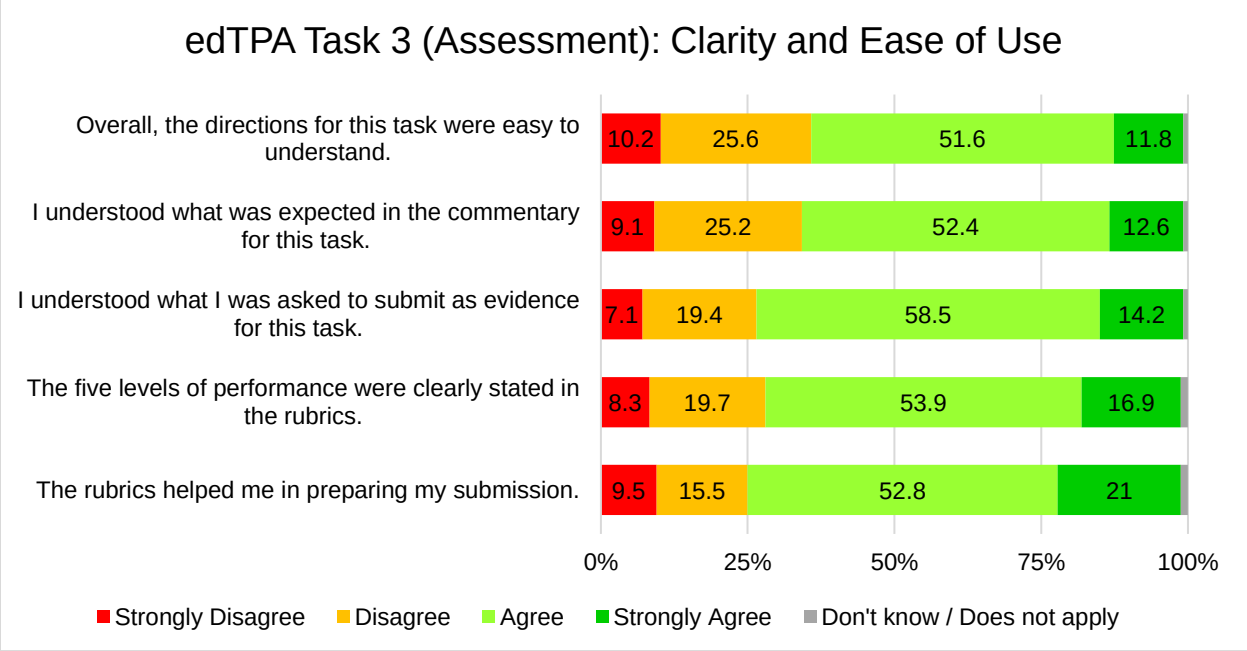

Figure 3.9. edTPA Task 3 (Assessment): Candidate perceptions of clarity and ease of use.

Next, candidates were asked to report their level of satisfaction with the edTPA resources available to them. The highest percentage of candidates (80.8%) thought that Understanding Rubric Level Progressions was a somewhat helpful or very helpful resource, followed by Making Good Choices (77.1%). Over 70 percent of candidates reported that the Academic Language Handouts were somewhat helpful or very helpful. Notably, 27.2 percent of candidates reported that the Candidate Registration Website was not helpful (see Figure 3.10).

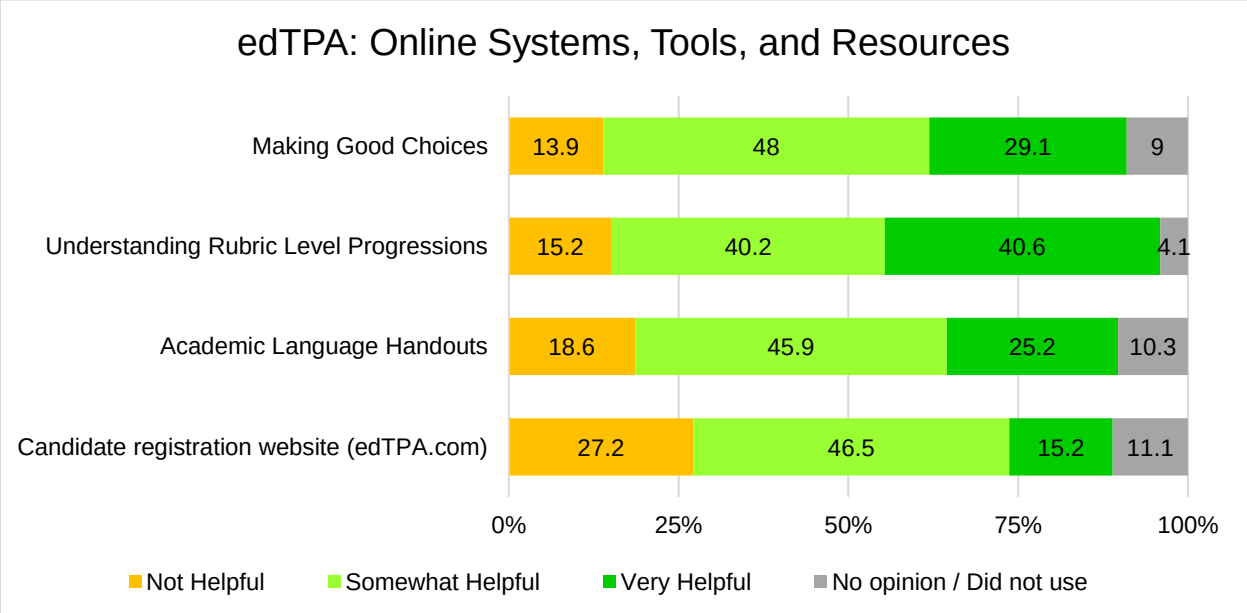

Figure 3.10. Candidate perceptions of edTPA resources.

Roughly 65 percent of candidates reported that they agreed or strongly agreed that the handbook and templates provided sufficient information to assist them throughout the assessment process (see Figure 3.11).

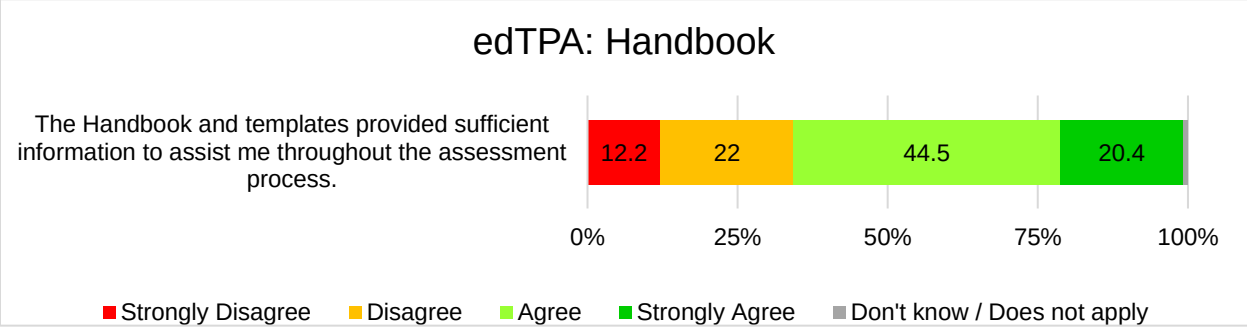

Figure 3.11. Candidate perceptions of edTPA handbook.

edTPA Survey Results for Claim 1 (Perceived Validity)

The results presented below in Figures 3.12, 3.13, and 3.14 address Claim 1; Claim 1 pertains to the face validity of edTPA (i.e., candidates' perceptions that the KSAs assessed by edTPA are emphasized in their program and that edTPA provides an opportunity to demonstrate those KSAs). The majority of respondents agreed or strongly agreed that Task 1 (66.7%), Task 2 (59.3%), and Task 3 (65.1%) provided them sufficient opportunity to demonstrate their knowledge, skills, and abilities related to the specific area of each of the Tasks. Approximately three out of every four candidate respondents reported that they agreed or strongly agreed that the KSAs assessed by the Tasks were emphasized in their preparation program.

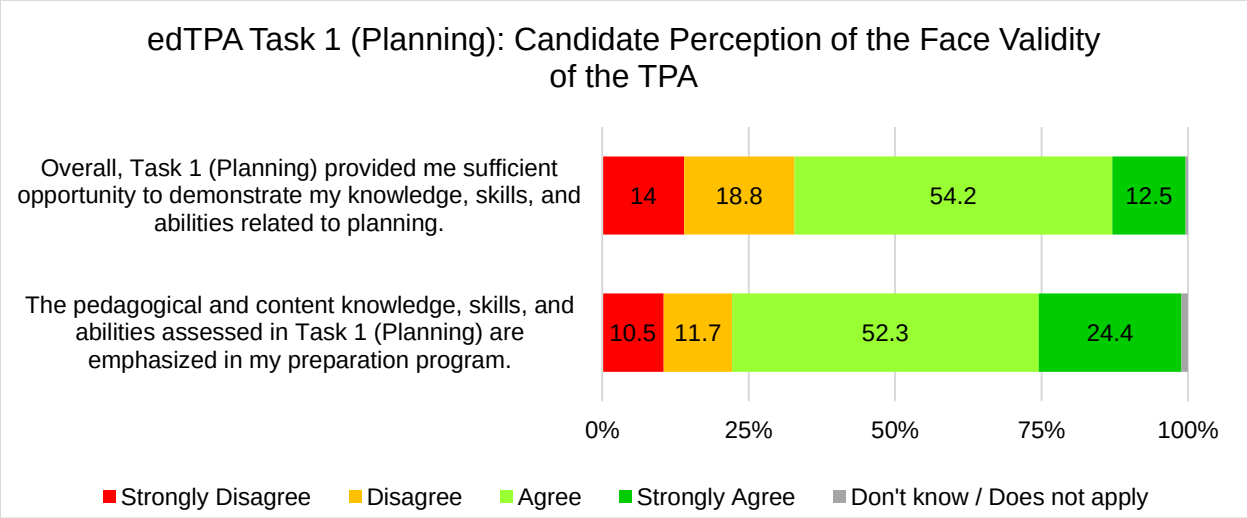

Figure 3.12. edTPA Task 1 (Planning): Candidate perceptions of validity.

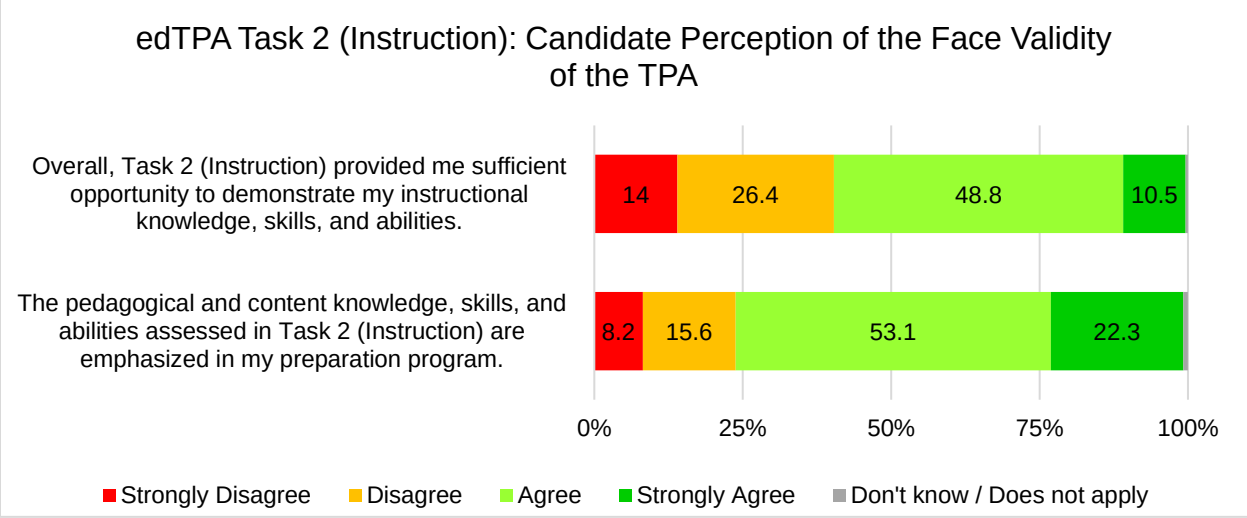

Figure 3.13. edTPA Task 2 (Instruction): Candidate perceptions of validity.

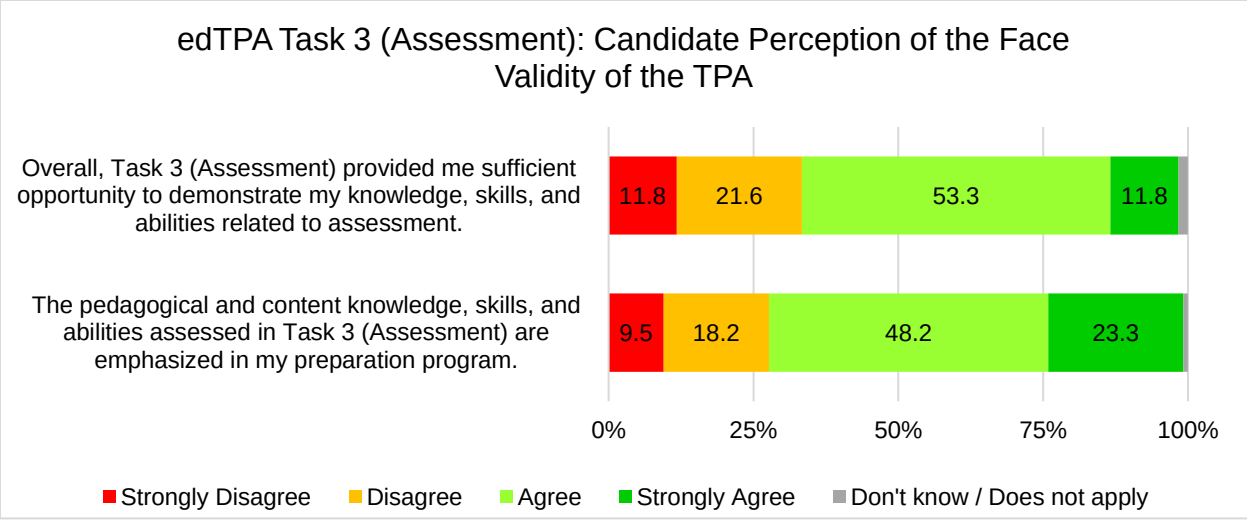

Figure 3.14. edTPA Task 3 (Assessment): Candidate perceptions of validity.

CalTPA Survey Results for Claim 2 (Clarity and Usefulness of Guidance/Supports)

Whereas the FAST and edTPA Candidate Surveys included the same survey items in Year 1 as in Year 2, the CalTPA model sponsors revised the Candidate Survey items for Year 2. Thus, there are some differences between the item-level results reported for the CalTPA Candidate Survey and the results reported for the FAST and edTPA Candidate Surveys.

Claim 2 pertains to the clarity and ease of use of the support materials/guidance provided to the CalTPA candidates by the model sponsor. Fifty-eight percent of CalTPA respondents reported that, “Overall, I had a clear understanding of CalTPA requirements.” Stated differently, roughly four out of 10 respondents disagreed or strongly disagreed that they had a clear understanding of CalTPA requirements (See Figure 3.15). Additional insight into this finding is provided in Figures 3.16 and 3.17.

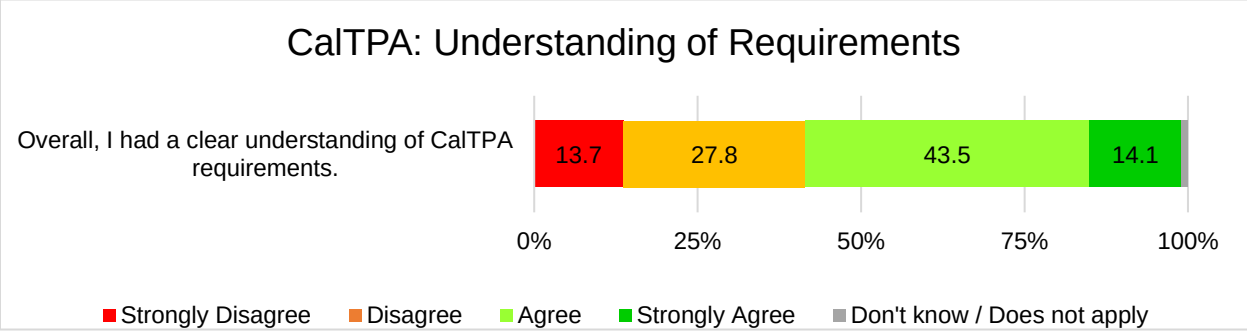

Figure 3.15. Candidate perceptions on the clarity of CalTPA requirements.

The CalTPA candidates were asked to evaluate the helpfulness of the resources provided to aid them with preparing and submitting their CalTPA submission. As shown in Figure 3.16, the majority of all candidates reported that the available resources were helpful, with the registration system (86.1%), the ePortfolio video annotation tool (83.6%), and the ePortfolio upload and submission system (82.5%) reported as the most helpful. The CalTPA support services (e.g.,

FAQs, Customer Support) were reported to be the least helpful and the resource of which candidates were least aware (see Figure 3.16).

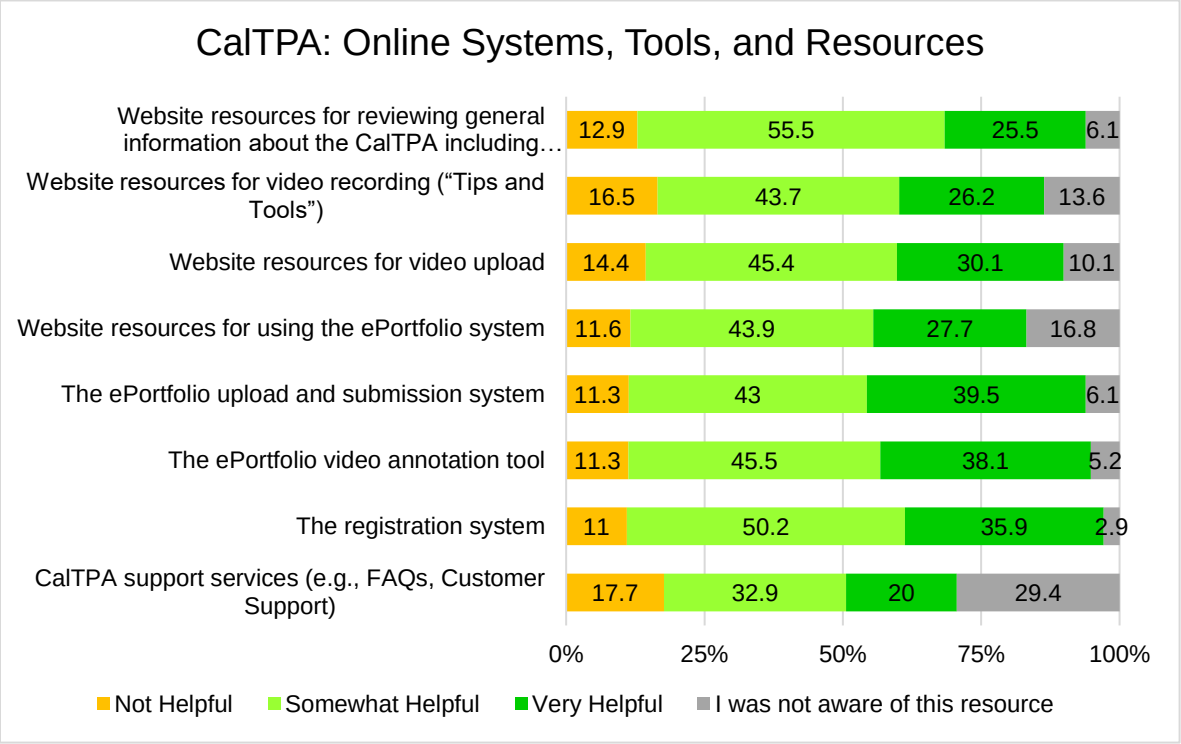

Figure 3.16. Candidate perceptions of CalTPA resources.

As shown in Figure 3.17, the majority of candidates reported that the Performance Assessment Guides were helpful (74.7%).

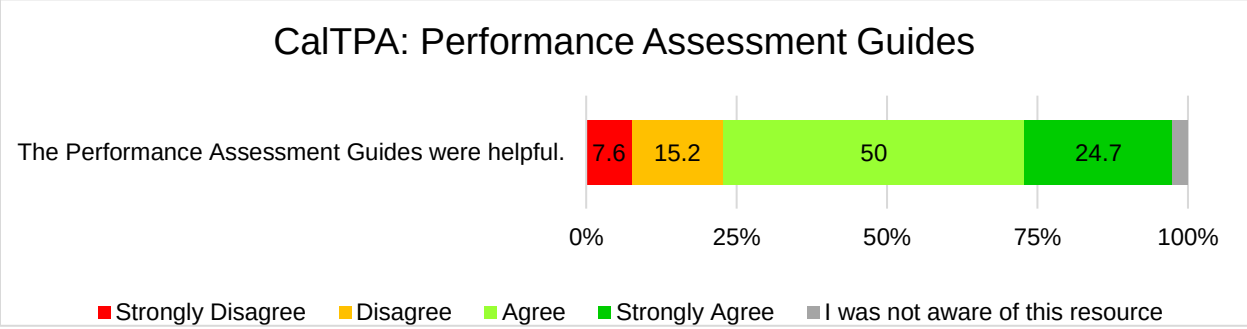

Figure 3.17. Candidate perceptions of CalTPA performance assessment guides.

CalTPA Survey Results for Claim 1 (Perceived Validity)

Data presented in Figure 3.18 address Claim 1 regarding the face validity of the CalTPA. The majority of candidates reported that the CalTPA assesses the knowledge, skills, and abilities emphasized in their preparation program (77.4%).

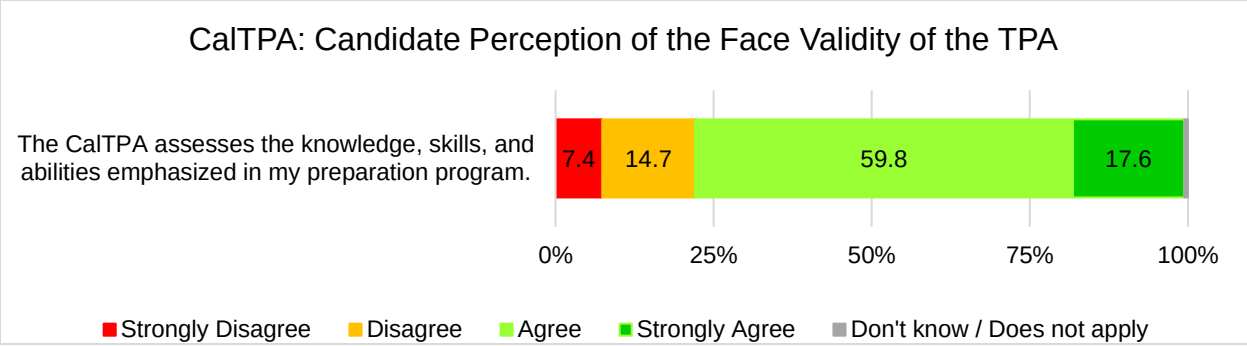

Figure 3.18. Candidate perceptions of validity.

Next, we present the findings from the Coordinator Surveys.

Coordinator Survey Results for Claim 2 (Clarity and Usefulness of Guidance/Supports)

Coordinator survey results are depicted across models within the same figures, given that the coordinators surveys did not have TPA component-specific items. The results presented below in Figure 3.19 address Claim 2, which pertains to how helpful the guidance/support materials available to the coordinators were in clarifying the purpose and requirements of the TPA. For each model, coordinators reported being clearer on the purposes of their specific TPA model than on the requirements of their specific TPA model. However, the majority of responses were still all very positive for FAST, edTPA, and CalTPA.

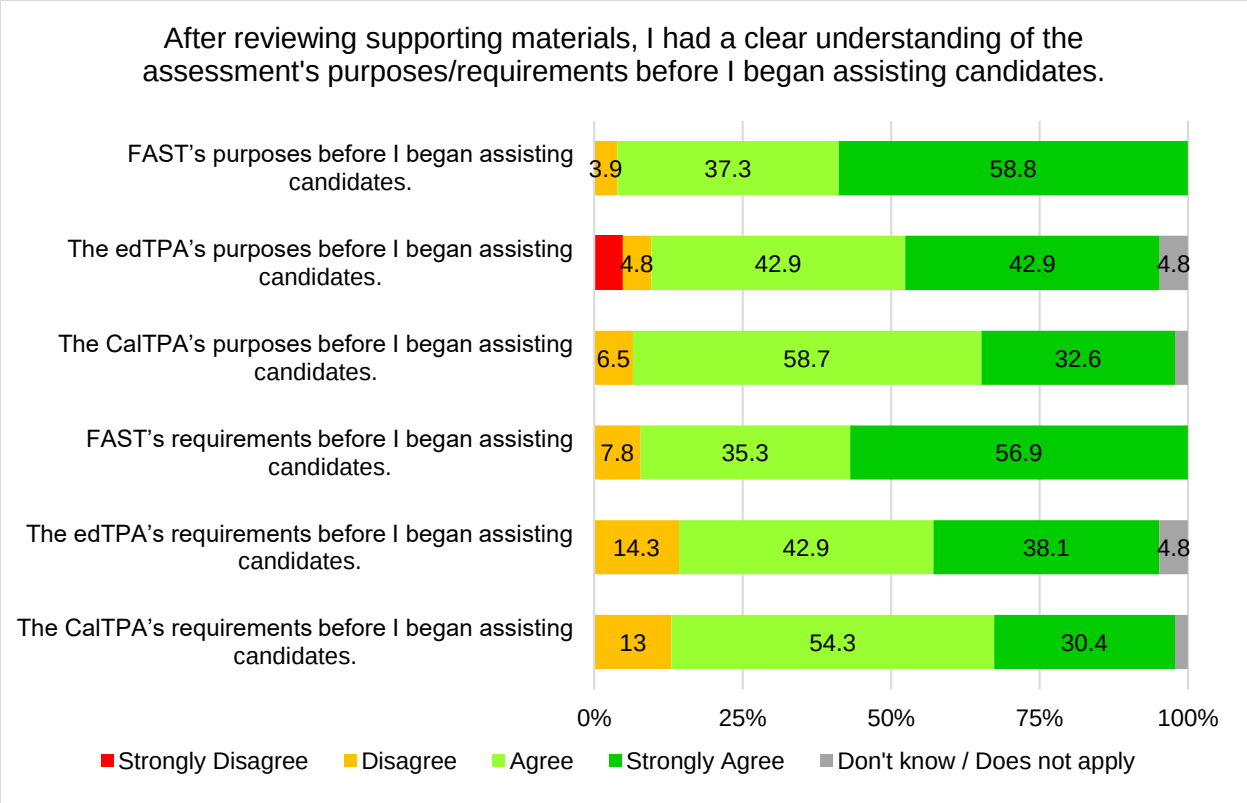

Figure 3.19. Coordinator perceptions of clarity.

As shown below in Figure 3.20, nearly 90 percent of respondents for each TPA model reported that they were well informed about their respective model during the process of assisting candidates. Coordinator respondents from FAST reported the highest levels of being informed with over 95 percent indicating that they agreed or strongly agreed they were well informed.

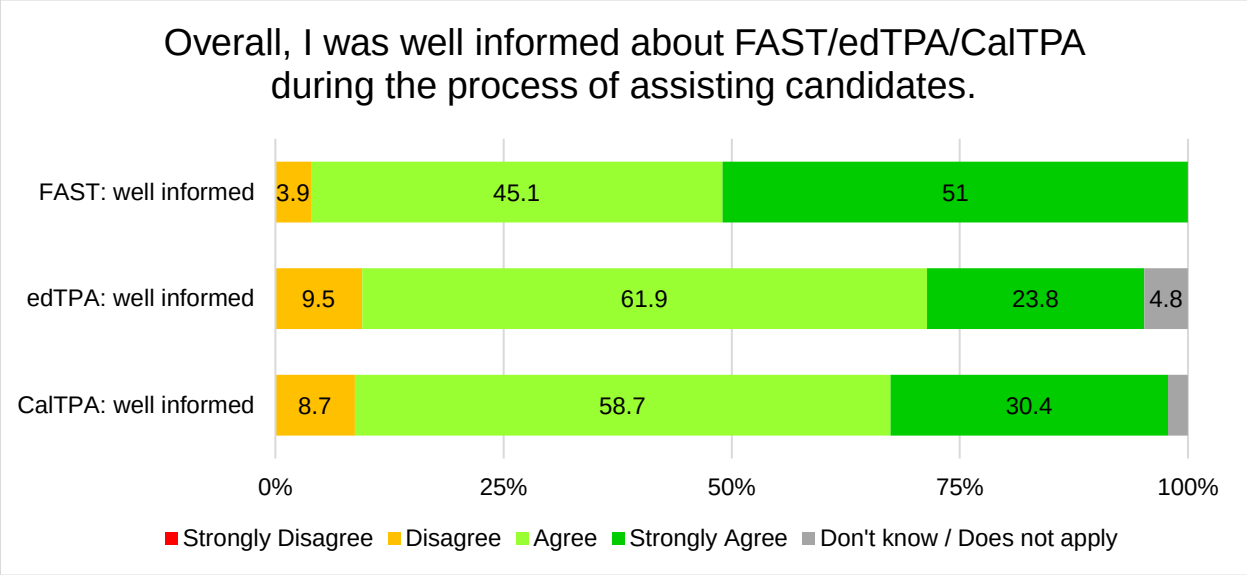

Figure 3.20. Coordinator perceptions of being well-informed.

In general, across all three models, coordinators had a positive view of the resources they used to access information about their TPA (see Figure 3.21). It's important to note, however, that each model uses different resources. Thus, Figure 3.21 is limited in the sense that it does not support direct one-to-one comparisons across models. Instead, Figure 3.21 simply portrays that, of the surveyed resources for each model, the majority of coordinators across all models positively endorsed the resources for their specific model.

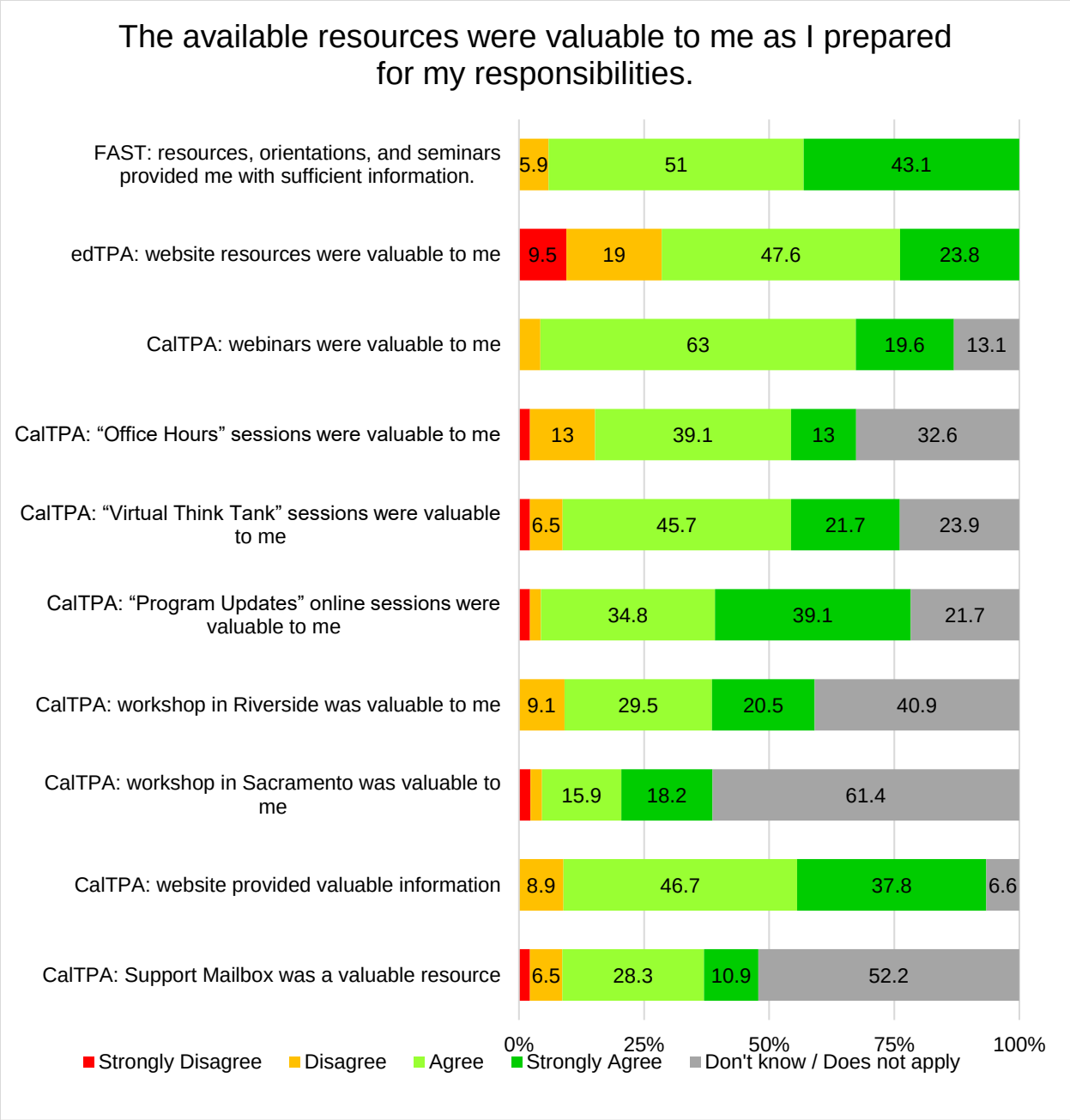

Figure 3.21. Coordinator perceptions of resources.

Coordinator Survey Results for Claim 1 (Perceived Validity)

The vast majority of coordinator respondents reported that they agreed or strongly agreed that their TPA focused on the appropriate skills and practices necessary for beginning teachers (see Figure 3.22). Coordinator respondents also reported that they agreed or strongly agreed that their TPA appropriately assessed candidate readiness in the areas measured, although roughly 20 percent of edTPA coordinators responded that they disagreed that their TPA appropriately assessed candidate readiness in the areas measured.

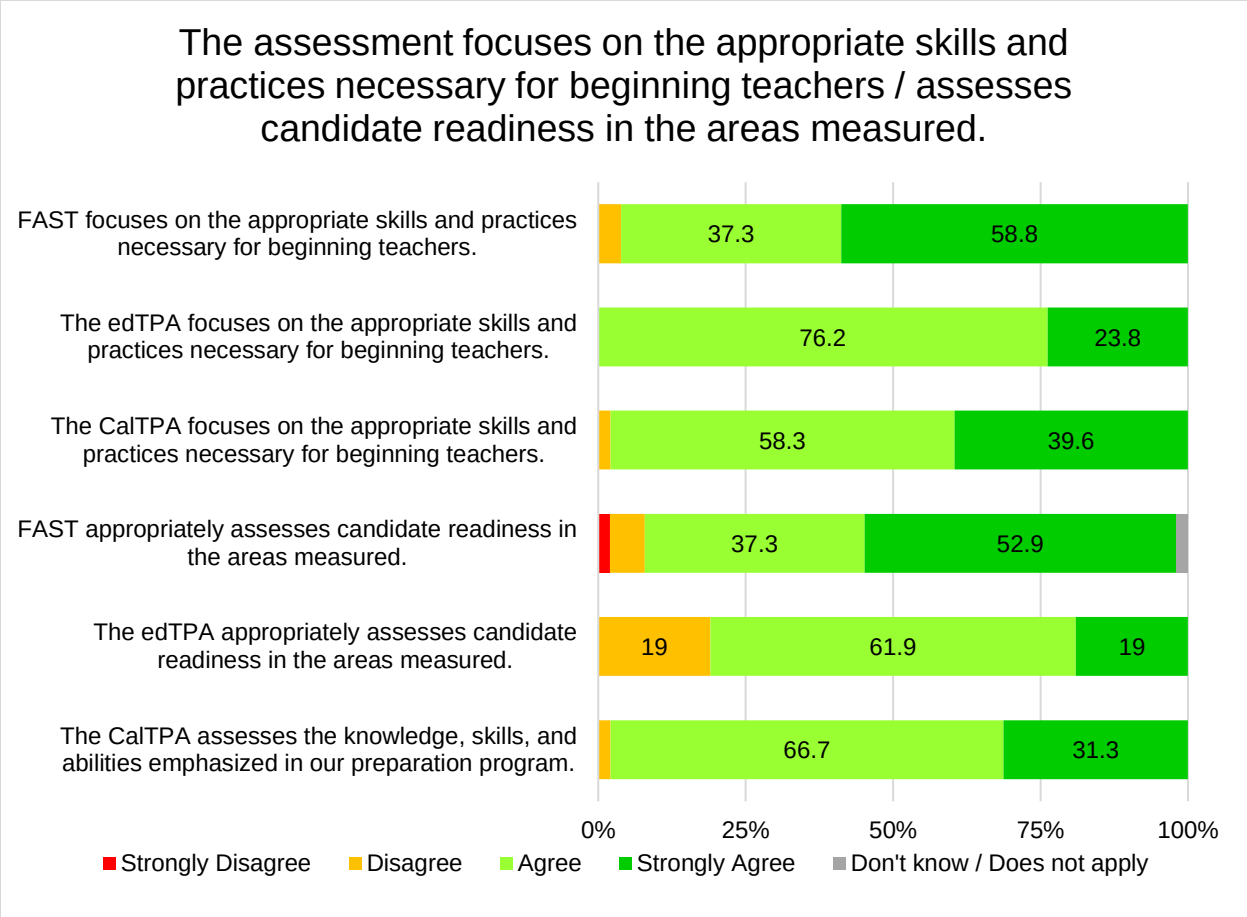

Figure 3.22. Coordinator perceptions of validity.

The Coordinator Surveys also asked coordinators to provide comments on their overall experience with the TPA. They were asked to be specific in their comments and include examples and reference specific steps and rubrics whenever possible. The themes that emerged from their comments are as follows: FAST coordinators commented that there have been issues with the online platform (TK20) and the portfolio submission process; edTPA coordinators reported that the amount of supporting documents is, at times, overwhelming; CalTPA coordinators reported that they would benefit from having TPA materials and information and access to the portal months ahead of time.

Discussion

The primary purpose of Activity 3 was to investigate Claim 2 by obtaining stakeholder perceptions of the clarity and usefulness of the guidance and supports provided to candidates and program coordinators by the model sponsor. This activity also helped to further inform Claim 1 by obtaining stakeholder perceptions about how well each model assesses the KSAs specified by the TPEs (i.e., the face validity of the TPAs).

Caveats and Limitations

The Candidate Surveys for FAST and edTPA are structured such that candidates are asked to respond to items about each component (task) of the TPA. The CalTPA Year 1 Candidate Survey, which served as the template for the FAST and edTPA surveys, was also structured this same way. For Year 2, the CalTPA model sponsors changed the structure of the CalTPA surveys. For this reason, direct comparisons in survey responses across models at the item-level is not always possible. Despite these item-level changes to the CalTPA surveys, the topics addressed by the Year 2 CalTPA surveys remain similar to the Year 1 topics. Consequently, some of the across model comparisons are provided at the topic level and other across model comparisons are provided at the item-level (i.e., for items on the CalTPA surveys that remained unchanged from Year 1 to Year 2).

Another limitation noted previously is the low response rate from edTPA candidates. Thus, results from the edTPA Candidate Survey, in particular, should be interpreted with caution.

Comparisons across TPA Models (Claim 2)

Candidates. Candidates for all three models were asked how clearly they understood the requirements for their TPA model (e.g., understanding the directions, rubrics, evidence requirements) based on the guidance and supports provided by the model sponsor. The majority of candidates agreed that they understood the requirements for their TPA model, although levels of agreement were strongest for the FAST candidates (80% or more), followed by edTPA candidates (59-75%, depending on the edTPA Task), and CalTPA (58%). Given that the FAST model is a local model developed and implemented only at Fresno State, it seems logical that candidates would have a strong understanding of their model. Also, given that CalTPA is not a locally developed model and was also in its first operational year, it seems logical that CalTPA candidates may not have had as clear of an understanding of model requirements as the locally developed FAST model or the well-established edTPA model.

All three Candidate Surveys also asked candidates about their model's Manual/Handbook/ Performance Assessment Guide. The majority of respondents for all three models indicated that this document provided sufficient information to assist them throughout the assessment process. For FAST and CalTPA, three out of four respondents agreed and for edTPA two out of three respondents agreed that these documents were helpful.

Each model's Candidate Survey also included survey items about the specific resources and supports available to candidates. Because edTPA and CalTPA are not local models and because they support a much larger candidate pool than Fresno State's FAST model, they provide more formal support materials to their candidates. The edTPA candidates reported the *Understanding Rubric Level Progressions* document to be among the most helpful resources provided to candidates. The CalTPA candidates reported the ePortfolio upload and submission system to be among the most helpful tools. On the other hand, roughly half the of the FAST candidates reported that the online system for uploading FAST submissions was *not* helpful.

Overall, these findings suggest that the guidance and supports provided by all three models are clear and useful for the majority of responding candidates and that the majority of candidates understand the requirements for their model. This should help to ensure that the models are implemented as intended. Apart from the continued challenges with the online system for uploading FAST submissions (which was also an issue in Year 1), the survey results suggest that the FAST candidates have the clearest understanding of their model's requirements. This

suggests that a small, localized model like FAST may be able to effect positive change more quickly than larger, decentralized models like CalTPA and edTPA.

Coordinators. The vast majority of responding coordinators across all three models reported that (a) after reviewing the support materials, they had a clear understanding of their model's purposes before they began assisting candidates, (b) they had a clear understanding of the requirements for their TPA model, and (c) they were well informed about their model. In terms of the resources available to coordinators, the majority of coordinators for all three models reported that the resources available to them (e.g., website resources, seminars/webinars) were valuable in helping them prepare for their responsibilities. Overall, these findings suggest that the guidance and support provided to coordinators are sufficiently clear and detailed, which helps to ensure that each model is implemented as designed and intended.

Comparisons across TPA Models (Claim 1)

Candidates. The survey results also help to address Claim 1. Claim 1 pertains to the perceived validity of the TPA models. The survey results indicate that the vast majority of candidates from all three models agree or strongly agree that the KSAs measured by their model are emphasized in their preparation program. This finding supports the validity of the TPA models from the perspective of candidates.

Coordinators. The survey results also lend support to Claim 1 from the perspective of the coordinators. Nearly all coordinators across all three models agreed that their TPA focuses on the appropriate skills and practices necessary for beginning teachers, and that their model appropriately assesses candidate readiness in the areas measured. If coordinators perceive that the TPAs are valid, then this should help to further ensure that coordinators are implementing the models as designed and intended.

Conclusion

The majority of candidates across all three models agreed that they understood the requirements (e.g., directions, rubrics, evidence requirements) for their TPA model, although understanding of requirements appeared to be strongest for FAST candidates. There was also consistently strong agreement across all models that coordinators had a clear understanding of their model's purpose and requirements, and that they felt well informed during the assessment process. Furthermore, the survey findings indicate that the TPA models are perceived as valid by both candidates and coordinators across all three models. These findings help to ensure that the TPA models are implemented as designed and intended, and thereby lend support to Claim 2 ("*The guidance and supports—e.g., guide/manual/handbook and other resources—provided by model sponsors to candidates and teacher preparation faculty are sufficiently clear and detailed to ensure that the model is implemented as designed and intended*.").

Chapter 4: Scoring Review – Comparison of Scoring Rubrics, Score Reports, and Rater Training (Activity 4)

Wade Buckland, Andrea Sinclair & Sunny Becker

Introduction

Activity 4 is an evaluation of the quality and comparability of edTPA, FAST, and CalTPA's scoring materials, practices, procedures, and validity evidence over the course of the 2017–18 and 2018–19 academic years. We investigated the following scoring-related claims (emphasis added):

- Claim 3: The **scoring rubrics** for each TPA model are sufficiently clear and detailed to ensure that trained raters can accurately and consistently score candidate submissions.
- Claim 4: For each TPA model, there is a comparable, comprehensive process to **select, establish calibration, and train** the **assessors** who score candidate submissions.
- Claim 7: For each TPA model, the **score reports** (candidate-level and program-level) provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used.
- Claim 8: The scoring rubrics and score reports provide **diagnostic information** on candidates and on programs such that the strengths and weaknesses of each can be identified.

This chapter is organized in four sections. First, the *Method* section details the procedures we used to conduct the review. Next, in *Results*, the three assessments are evaluated separately based on the scoring materials and procedure criteria we developed. Then, in our *Discussion* section, we discuss the evidence across all three assessments for each claim in terms of the relevant evaluation criteria. Finally, the *Conclusion* section provides summary remarks.

Method

Our methodology consisted of three tasks: a) Scoring Procedure Document Review and Data Collection, b) Scoring Evaluation Criteria Development, and, c) Scoring Evaluation.

Scoring Procedure Document Review and Data Collection

1. HumRRO researchers assembled and collected documentation and materials relevant to Claims 3, 4, 7, and 8. Documents were collected from a) HumRRO's file library created to complete Activity 1 (see Chapter 1), and b) additional materials collected from model sponsors specific to scoring. It was particularly important to collect revised documentation and materials from CalTPA model sponsors to reflect revisions to CalTPA in 2018–19 (first operational year) that stemmed from information learned during the 2017–18 field test.
2. To supplement our documentation review, a pair of HumRRO researchers with substantial expertise in the evaluation of rubrics, score reports, and human scorer training procedures observed in-person assessor training and calibration for FAST and CalTPA. Because edTPA had no in-person training or calibration events to attend, we completed edTPA's asynchronous, on-line training hosted on Pearson's ePEN software platform. During our visits, we completed checklists (See Appendices 4.A and 4.B for

these forms) to structure our notes.[25] We also conducted short interviews with trainers and trainees during breaks; these informal interviews provided insight into processes that might not be readily observable (e.g., how benchmarks/markers were selected for calibrating raters). Table 4.1 presents a listing of the title, date(s), mode (i.e., webinar or onsite), and location of our observations.

Table 4.1. Scoring Training and Calibration Observations

Scoring Training Event	Date(s)	Mode / Location
FAST		
Mathematics TSP Assessor Training and Calibration	4/9/18	Onsite/Fresno, CA
Multiple Subject TSP Assessor Training & Calibration	4/9/18	Onsite/Fresno, CA
edTPA		
Online Training	May 2018	Remote/Asynchronous
Calibration	None	None
CalTPA		
Assessor Orientation	1/19/18	Webinar
Assessor Orientation	2/13/18	Webinar
Cycle 1 Assessor Training & Calibration (southern)	3/20/18	Onsite/San Bernardino, CA
Cycle 2 Assessor Training & Calibration (northern)	4/11/18	Onsite/Sacramento, CA
Cycle 1 & 2 Assessor Training & Calibration (southern)	11/26-27/18	Pomona, CA

3. Between December 2018 and April 2019, we conducted telephone interviews with TPA model sponsors to ask specific questions about scoring procedures and processes. These interviews were intended to ensure we had a complete understanding of each model's scoring methods, documentation, and 2018–19 updates and changes. April interviews were recorded and transcribed, and notes were taken during interviews. Two HumRRO staff members participated in each call. We conducted interviews with:
    a. Amy Reising, CalTPA's Director of Performance Assessment Development on December 14, 2018,
    b. Helene Mandell, a CalTPA trainer and recent Director of Field Experiences at University of San Diego on April 24, 2019,
    c. Jeanie Behrend, FAST Coordinator on April 25, 2019, and
    d. Nicole Merino, edTPA's Director of Performance Assessment on April 26, 2019.

Scoring Evaluation Criteria Development

1. A pair of HumRRO staff members independently reviewed the *Assessment Design Standards* (ADS) to map relevant ADS to Claims 3, 4, 7, and 8. Then, they met to adjudicate any differences in the ADS they selected as relevant to Claims 3, 4, 7 and 8.

2. The same researchers independently reviewed the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education's *Standards for Educational and Psychological Testing* (2014)—hereafter, *Joint Standards*—to identify relevant *Joint Standards* for Claims 3, 4, 7, or 8.[26]

[25] Appendices for this report are in Volume II: Appendices.
[26] We capitalize "Standard" throughout this chapter when referring to a standard specified by ADS or the *Joint Standards*, as opposed to a standard that is a generally accepted expectation in the industry.

3. After conducting Step 2, the researchers compared selected *Joint Standards* to adjudicate differences in the relevant *Joint Standards* identified. Unlike Activity 1 (see Chapter 1), *Joint Standards* were selected from all chapters—not just the Test Design and Development chapter. The additional *Joint Standards* were included in Activity 4 to ensure all validity, reliability, scoring, and interpretation Standards that are relevant to Claims 3, 4, 7, and 8 were evaluated.

4. The researchers combined and organized the selected Standards from these two sources (i.e., the ADS and *Joint Standards)* by claim. In some cases, more than one claim aligned with each Standard.

5. We isolated statements from the selected Standards to facilitate comparisons across TPA models, reduce repetition, better align the Standards with our claims, and ensure a comprehensive evaluation. To illustrate this process, below we present *Joint Standard* 4.18 (with comment) and the derived statements we isolated from it. For a complete list of the full unedited ADS and *Joint Standards* that we found to be aligned to the claims, see Appendix 4.C.

> *Joint Standard* 4.18: Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.
>
> Comment: In scoring more complex responses, test developers must provide detailed rubrics and training in their use. Providing multiple examples of responses at each score level for use in training scorers and monitoring scoring consistency is also common practice, although these are typically added to scoring specifications during item development and tryouts. For monitoring scoring effectiveness, consistency criteria for qualifying scorers should be specified, as appropriate, along with procedures, such as double-scoring of some or all responses. As appropriate, test developers should specify selection criteria for scorers and procedures for training, qualifying, and monitoring scorers. If different groups of scorers are used with different administrations, procedures for checking the comparability of scores generated by the different groups should be specified and implemented.

HumRRO Evaluative Statements Derived from *Joint Standard* 4.18.

- Procedures for [scoring] are presented by the [model sponsor] with sufficient detail and clarity to maximize the accuracy of scoring.
- [Scoring criteria are] presented by the [model sponsor] with sufficient detail and clarity to maximize the accuracy of scoring.
- Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses are clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.
- [The model sponsor provides] multiple examples of responses at each score level for use in training scorers and monitoring scoring consistency. [These] are typically added to scoring specifications during item development and tryouts.

- [The model sponsor specifies] consistency criteria for qualifying scorers.
- [The model sponsor specifies] consistency criteria for double-scoring of some or all responses.
- [The model sponsor specifies] selection criteria for scorers.
- [The model sponsor specifies] procedures for training, qualifying, and monitoring scorers.
- [The model sponsor specifies and implements] procedures for checking the comparability of scores generated by the different groups of scorers if different groups are used with different administrations.

6. After creating the final list of Standards (or criteria/isolated statements) aligned with the claims, HumRRO's project director conducted a cross-check. The finalized criteria are presented in the main body of this report.

Scoring Evaluation

1. In the spring and summer of 2018 (i.e., Year 1 of the comparability study), two HumRRO researchers independently rated each TPA model on the finalized list of standards-based statements. For each rating, a rationale was provided. Both staff members were trained and calibrated on the rating scale. Discrepancies were discussed until the raters came to consensus.

2. In Year 2, HumRRO researchers (a) collected documentation related to changes that occurred to the TPA models for the 2018–19 year, (b) located previously unknown information from Year 1, and (c) conducted interviews with model sponsors in December 2018 - April 2019. Then, the two HumRRO researchers updated ratings on the standards-based statements and their associated rationales. Discrepancies were again discussed until the raters came to consensus.

3. The draft report was submitted to the TAC for review in June 2019 and feedback was provided. HumRRO revised the ratings based on the TAC feedback.

Table 4.2. Rating Scale for Strength of Evidence

Rating Level	Description of Rating Levels
1	No evidence of the Standard/element found in the documentation provided.
2	Little evidence of the Standard/element found in the documentation; less than half of the Standard/element covered in the documentation and/or evidence of key aspects of the Standard/element could not be found.
3	Some evidence of the Standard/element found in the documentation; approximately half of the Standard/element covered in the documentation including some key aspects of the Standard/element.
4	Evidence in the documentation mostly covers the Standard/element; more than half of the Standard/element covered in the documentation, including key aspects of the Standard/element.
5	Evidence in the documentation fully covers all aspects of the Standard/element.

Results

HumRRO researchers agreed that seven elements, 1a, 1h, 1g, 2a, 2c, 2e, and 2g, from the ADS aligned with Claims 3, 4, 7, and 8. Three of the elements, 1a, 1h, and 1g, were derived from *ADS 1: Assessment Designed for Validity and Fairness*. Four elements, 2a, 2c, 2e, and 2g, were derived from *ADS 2: Assessment Designed for Reliability and Fairness.* Table 4.3 displays the seven ADS elements and 12 *Joint Standards* we identified as aligned to Claims 3, 4, 7, and 8.

Table 4.3. Assessment Design Standard Elements and Joint Standards Aligned to Claims 3, 4, 7, and 8

Claim	Assessment Design Standard	Joint Standard
3	1(a), 1(c), 1(h)	4.18
4	1(g), 1(h), 2(c), 2(e)	3.0, 3.4, 3.8, 4.20, 6.9
7	2(g)	1.1, 1.2, 2.13, 3.8, 4.22, 5.0, 6.10
8	1(a), 2(a)	1.1, 6.10

Table 4.4 displays the number of evaluative statements, aligned to Claims 3, 4, 7, and 8, that we derived from each identified ADS element and *Joint Standard*. Again, evaluative statements are discrete, derived statements isolated from the lengthier Standard. An example of an evaluative statement derived from *Joint Standard* 4.18 is, "*Procedures for [scoring] are presented by the [model sponsor] with sufficient detail and clarity to maximize the accuracy of scoring*." For each ADS, we derived 1 to 4 statements. For each *Joint Standard*, we derived 1 to 6 statements. In total, we derived 49 statements with 19 statements from the ADS and 30 statements from the *Joint Standards*.

Table 4.4. Number of Evaluative Statements Derived from each Standard by Claim

Standard	Claim 3	Claim 4	Claim 7	Claim 8	Total
Assessment Design Standard					
1(a)	1			1	2
1(c)	1				1
1(g)		1			1
1(h)	2	1			3
2(a)				1	1
2(c)		4			4
2(e)		4			4
2(g)			3		3
ASD Total	4	10	3	2	19

(continued)

Table 4.4. (Continued)

Standard	Claim 3	Claim 4	Claim 7	Claim 8	Total
Joint Standard					
1.1			2	1	2
1.2			1		1
2.13			1		1
3.0		1			1
3.4		1			1
3.8		2	1		3
4.18	3				3
4.20		4			4
4.22			1		1
5.0			2		2
6.9		6			6
6.10			1	3	4
JS Total	3	14	9	4	30
Grand Total	**7**	**24**	**12**	**6**	**49**

Next, for each assessment, we present the results of the numeric ratings assigned to each evaluative statement. If evidence for all or part of a Standard was available for review, we provided a rating (using the rating scale presented in Table 4.2). The results are presented in Tables 4.5 – 4.8 for FAST, Tables 4.9 – 4.12 for edTPA, and Tables 4.13 – 4.16 for CalTPA, respectively. Each table includes the (a) number for the Standard/isolated evaluative statement in the left column, (b) Standard/isolated evaluative statement in the left-middle column, (c) rating on the strength of evidence for the Standard in the middle-right column, and (d) rationale for the rating in the far-right column.

FAST Results

Table 4.5. Claim 3 Ratings on the Assessment Design and Joint Standards for FAST

#	Assessment Design Standard/Joint Standard	FAST Rating	Rationale for FAST Rating
3.1	ADS 1(a) The assessment [includes] multi-level scoring rubrics that are clearly related to the TPEs that the task measures.	5	FAST uses four-point rubrics for both the Site Visitation Project (SVP) and Teaching Sample Project (TSP). Per the FAST Tasks Matrix, the sponsor has documented the relationship between the TPE elements (i.e., key aspects) and tasks in the FAST Tasks Matrix. Each component of the two FAST tasks are linked to at least two major domains of the TPEs. Moreover, the FAST Manual demonstrates how the FAST rubrics align with the TPE elements. The TSP score report also maps rubric scores to TPE elements.
3.2	ADS 1(c) The model sponsor defines scoring rubrics so candidates for credentials can earn acceptable scores on the Teaching Performance Assessment with the use of different content-specific pedagogical practices that support implementation of the TK-12 content standards and curriculum frameworks.	5	SVP and TSP rubrics are neutral with regard to subject-specific pedagogical practices. Language within the rubrics is general enough to allow candidates to earn acceptable scores with the use of different subject-specific pedagogical practices and curriculum frameworks. The FAST Manual and TSP Score Report show how the FAST rubrics are mapped to the TPE elements, which are directly and purposely aligned to the TK-12 content standards and curriculum frameworks.
3.3	ADS 1(h) The model sponsor develops scoring rubrics that focus primarily on teaching performance.	5	The three SVP rubrics and seven TSP rubrics in the FAST Manual (2.0) are aligned with the TPEs, which are primarily focused on teaching performance (see “TPE Elements Assessed by FAST 2.0 Rubrics” in Table 1 of the FAST Manual). Furthermore, after the 2017–18 school year (field test), FAST made changes to the wording of the rubrics for the 2018–19 school year (operational) to improve clarity. These revisions stemmed from feedback provided by coaches and teacher candidates (interview with model sponsor, April 27, 2019).
3.4	ADS 1(h) The model sponsor develops scoring rubrics 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	FAST’s scoring rubrics only relate to pedagogical competence. Other factors such as personal attire, appearance, demeanor, speech patterns and accents are not evaluated. An additional safeguard to ensure that scoring rubrics focus on teaching performance is anti-bias scorer training to reduce the effect of non-pedagogical performance factors on scores.

(continued)

Table 4.5. (Continued)

#	Assessment Design Standard/Joint Standard	FAST Rating	Rationale for FAST Rating
3.5	JS 4.18 Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.	3	Oral instructions for using the rating scales were provided to FAST assessors at the observed April 2018, 2-hour TSP calibration training sessions. Scorer training focused on teaching assessors to look for evidence to support scores. During the observed assessor training, the pair of trainers—who trained each credential area—told assessors to evaluate each submission on each indicator of each scoring rubric (e.g., evaluate both Implications for Instruction and Creating and Maintaining Effective Environments indicators within the Students in Context rubric). If both indicators were at the same level, assessors were instructed to simply use that level as the overall rating for the rubric. However, if the indicators were at different levels, assessors were instructed to determine the level of the overall rubric rating based on the assessor's judgment. While the FAST model sponsor indicated that most indicators are weighted evenly [interview on April 27, 2019], we recommend that a written guideline for instructing assessors how to weight the importance of individual indicators for the overall rubric rating should be added to the "FAST 2.0 Scorer Training Procedures" document. Indicators like "Reflection" have three indicators and specific indicators may need to be weighted differently. If different assessors are inconsistently weighting indicators within a rubric in assigning scores, then this could threaten the inter-rater reliability of the scores.
3.6	JS 4.18 [Scoring criteria are] presented by the [model sponsor] with sufficient detail and clarity to maximize the accuracy of scoring.	3	See rationale for 3.5 above.
3.7	JS 4.18 [The model sponsor provides] multiple examples of responses at each score level for use in training scorers and monitoring scoring consistency. [These] are typically added to scoring specifications during item development and tryouts.	3	During training, assessors review two exemplars/benchmarks, representing two full individual candidate submissions, as a group activity. Then, they independently score a third exemplar for calibration. The example submissions (exemplars) do not show examples of responses at every score point. More examples of score points for each indicator/rubric for each subject area should be provided to assessors as scoring benchmarks.

Table 4.6. Claim 4 Ratings on the Assessment Design and Joint Standards for FAST

#	Assessment Design Standard/Joint Standard	FAST Rating	Rationale for FAST Rating
4.1	ADS 1(g) The TPA model sponsor [provides] materials appropriate for use by [assessors] to become familiar with the design of the TPA model, the candidate tasks, the scoring rubrics, [and scoring processes].	5	The FAST model sponsor provided assessors with the FAST Manual, which includes a detailed explanation of (a) FAST's format, (b) instructions for completing the tasks, (c) scoring rubrics, (d) TPE element to Task (SVP and TSP) alignment, and (e) policies and procedures. Also, the model sponsor meets with assessors at the beginning of every semester to remind them of and talk through the scoring process. The "FAST 2.0 Scorer Training Procedures" appendix of the FAST Response to the Assessment Design Standards document provides an overview of each task; assessor guidelines; bias training; training steps; and calibration/qualifying steps.
4.2	ADS 1(h) The model sponsor develops 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	A component of the FAST assessor training is the Personal Bias Hit List activity based on training materials developed by Idaho State University. This involves writing out personal biases and sharing them with the group. It is emphasized that candidates should not be rated on factors such as personal attire, appearance, demeanor, speech patterns and accents. Further, the FAST representative emphasizes that writing ability, grammar, and spelling are not part of the TPEs. At 2018 TSP Training, assessors were cautioned to not let poor skills in these areas bias their ratings. This activity was continued in the 2018–19 assessor training and is included in "FAST 2.0 Scorer Training Procedures."
4.3	ADS 2(c) The assessor training program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and the multi-level scoring rubrics.	5	FAST uses its faculty to score candidate submissions. Because of this, one can infer that they have a deep understanding of the TPEs and pedagogical assessment tasks. Furthermore, FAST rubrics are mapped to the TPEs. So, as assessors are gaining deep understanding of the rubrics, they're also increasing their understanding of the TPEs. At the spring 2018 training, researchers observed that each rubric for each task is read and discussed at length. The "FAST 2.0 Scorer Training Procedures" document also demonstrates that assessors review and discuss TPEs, tasks, and scoring rubrics during training. Training is differentiated for New Scorers and Experienced Scorers. New Scorer Training includes in-depth exploration of TPEs, task directions and rubrics, discussion of candidate work, and discussion of scores. The FAST rubrics are mapped to the TPE elements. So, as assessors are gaining deep understanding of the rubrics, they're also increasing their understanding of the TPEs. The model sponsor also meets with assessors (coaches) at the beginning of each semester to remind them of and talk through the scoring process.

(continued)

Table 4.6. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
4.4	ADS 2(c) The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy and calibration in relation to the scoring rubrics associated with the task.	4	FAST assessors are required to attend scorer training before scoring teacher candidate submissions. Training includes calibration on subject-specific candidate submissions with a trainer. The Spring 2018 FAST training sessions for TSP took place at separate times for different subject-specific pedagogy groups. During the observed Mathematics and Multiple Subjects training sessions, calibration was conducted in groups (with four trainees at the Mathematics session and two trainees at the Multiple Subjects session). During the sessions, the trainees and trainer(s) discussed why a response earned a specific point value on each rubric for two candidate submissions from the trainees' subject area. Independent ratings by trainees for calibration were not collected or evaluated for field test. In 2018–19, calibration materials and procedures improved from field test (2017–18) because exemplars for calibration were available for the 2018–19 assessor training. Training procedures also improved in 2018–19 because after group training on two exemplars, assessors were required to independently score a third exemplar on which they had to meet a performance threshold; to meet the calibration threshold scorers needed exact matches on at least four of the seven rubrics. Scores that were not exact matches had to be at least one score point adjacent to the correct score point. Returning scorers were not required to re-calibrate in 2018–19 despite changes to rubrics in 2018–19, although returning scorers did attend training in which edits to the rubrics were discussed. Scorer performance data on calibration exercises are unavailable, but all participants qualified to score [interview with model sponsor, April 27, 2019].

(continued)

Table 4.6. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
4.5	ADS 2(c) The model sponsor uses only assessors who successfully calibrate during the required TPA model assessor training sequence.	3	As of 2018–19, scorers were required to meet a performance standard or were required to remediate scoring deficiencies before becoming calibrated. During the calibration exercise, scorers needed exact matches for all rubrics except three. Scores that were not exact matches had to be at least one score point adjacent to the correct score point. Scorer performance data on calibration exercises are unavailable but all participants qualified to score [interview with model sponsor, April 27, 2019]. The calibration is used as a formative evaluation to determine how to remediate scorers who do not calibrate on their first try. The rubrics changed from the field test year (2017–18) to the operational year (2018–19). The model sponsor described these changes as "minor" and "no substantive changes." However, nearly all descriptors were revised, and revisions included (a) wording changes, (b) the addition of criteria, and (c) the deletion of criteria. See Appendices 4.D (SVP) and 4.E (TSP) in Volume II of this report for a sample of the comparisons between field test rubrics and operational rubrics. An example of the change from field to operational is the Level 2 rating description for Learning Outcomes and Standards within the Learning Outcomes rubric for the TSP: In 2017–18 (field test) the description read: *Standards and outcomes primarily address either content knowledge or literacy skills. Some outcomes represent the content and level of learning reflected in the content standards. Outcomes reflect a limited range in the type of level of learning.* In 2018–19 (operational), the description was updated to read: *Outcomes primarily address either content or literacy standards. Most outcomes represent the content and level of learning (e.g., DOK level) reflected in the content standards, though they primarily focus on lower levels of learning.* Returning scorers were not required to re-calibrate in 2018–19 on the revised rubrics [interview with model sponsor, April 27, 2019], although returning scorers did attend review sessions to discuss changes to the rubric and to discuss exemplars in areas that seemed most problematic for candidates and scorers.[27] When such changes are made to rubrics, re-calibration of assessors is strongly recommended.
4.6	ADS 2(c) When new pedagogical tasks and scoring rubrics are incorporated into the assessment, the model sponsor provides additional training to the assessors, as needed.	5	FAST piloted its assessment in 2016–17 and field tested it in 2017–18. New training was provided in 2017–18 and again in 2018–19. All scorers attended scorer training, including returning scorers (although returning scorers were not required to re-calibrate even though changes were made to the rubrics from 2017–18 to 2018–19).

(continued)

[27] Clarification provided by model sponsor on 13 July 2019 in response to review of draft report.

Table 4.6. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
4.7	ADS 2(e) All approved models must include a local scoring option in which the assessors of candidate responses are program faculty and/or other individuals identified by the program who meet the model sponsor's assessor selection criteria. These local assessors are trained and calibrated by the model sponsor, and whose scoring work is facilitated, and their scoring results are facilitated and reviewed by the model sponsor.	5	FAST is a locally developed model. Thus, all scoring is local and is conducted by trained and calibrated faculty.
4.8	ADS 2(e) The model sponsor must provide an annual audit process that documents that local scoring outcomes are consistent and reliable within the model for candidates across the range of programs using local scoring and informs the Commission where inconsistencies in local scoring outcomes are identified. If inconsistencies are identified, the sponsor must provide a plan to the CTC for how it will address and resolve the scoring inconsistencies both for the current scoring results and for future scoring of the TPA.	5	Appendix G in the Response to the Assessment Design Standards document (provided to the Commission in May 2018) includes statistical analyses that demonstrate adherence to the designated thresholds for consistency/reliability of double scored submissions for SVP and TSP. Analyses are collapsed across multiple subject and single subject credential areas. Future analyses should report agreement rates by credential area when there are sufficient numbers of submissions to do so (i.e., approximately 30); this will help to ensure that scoring outcomes are reliable across programs and credential areas (and not just for some).
4.9	ADS 2(e) The model sponsor provides a detailed plan for establishing and maintaining scorer accuracy and inter-rater reliability during field testing and operational administration of the assessment.	3	The model sponsor provides a plan for *establishing* scorer accuracy and inter-rater-reliability during field testing and operational administration. However, at present, double scoring is used only to report on inter-rater reliability *after* candidates receive their score of record. Plans should be made to ensure submissions are double scored throughout the scoring window and that feedback/retraining is provided to scorers with low reliability *during* the scoring window. This is key to *maintaining* scoring accuracy and preventing scorer drift.
4.10	ADS 2(e) The scoring process conducted by the model sponsor to assure the reliability and validity of candidate outcomes on the assessment may include, for example, regular auditing, selective back reading, and double scoring of candidate responses near the cut score by the qualified, calibrated scorers trained by the model sponsor.	4	Approximately 15% of submissions are double scored. Double scoring is used to report on reliability/consistency of scorers. Interrater reliability analyses from 2017–18 show that scorers were reliable. Double scoring is not used to provide feedback to low reliability scorers as a way to increase their scoring accuracy during the scoring window. See also rationale for 4.9. Per the documentation in the Response to Assessment Design Standards (May, 2018), "any task that receives a non-passing score on any rubric section is re-scored by a second trained scorer. If there is a discrepancy between the two scores, a third scorer will be used to determine the score (p. 20)." This helps to assure the reliability and validity of candidate outcomes on the assessment.

(continued)

Table 4.6. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
4.11	JS 3.0 All [scoring procedures steps] are designed in such a manner as to minimize construct-irrelevant variance.	3	Double scoring should occur during the scoring window so that scorers with low reliability (which could be indicative of a rating bias) can be identified and re-trained.
4.12	JS 3.4 Test takers receive comparable treatment during the [scoring process]. Those responsible for testing adhere to standardized scoring protocols so that test scores will reflect the construct(s) being assessed and will not be unduly influenced by idiosyncrasies in the testing process.	3	Use of few scorers per subject area (within the Single Subject program) without monitoring throughout the scoring window and scorer-candidate familiarity in the Single-Subject Credential areas are threats to comparable treatment of teacher candidates during the scoring process.
4.13	JS 3.8 Adequate training and calibration of scorers is carried out and monitored throughout the scoring process to support the consistency of scorers' ratings for individuals from relevant subgroups. Where sample sizes permit, the precision and accuracy of scores for relevant subgroups also is calculated.	3	FAST training and calibration are described in the rationales for 4.1 to 4.5 and monitoring of scorers is described in the rationales for 4.9 to 4.11. Calibration of scorers is not monitored throughout the scoring window. Double scoring occurs *after* the scoring window has ended and is used for documentation purposes only, not to provide feedback to low reliability scorers during the scoring window. After scoring ends, FAST compares ethnicity, self-rated English language fluency, self-reported disability, and gender category subgroup subtest and final score performance after each administration using Mann-Whitney *U* and Kruskal-Wallis *H*. Results can be found in Appendix G of the FAST Response to the Assessment Design Standards (May, 2018). Small sample sizes of some subgroups (e.g., Limited Working Proficiency in English, Asian, Black) should be accumulated over time to build a large of enough sample to conduct these subgroup analyses.
4.14	JS 3.8 For human scoring, scoring procedures [are] designed with the intent that the scores reflect the examinee's standing relative to the tested construct(s) and are not influenced by the perceptions and personal predispositions of the scorers.	4	While FAST assessor training includes personal bias training and discussion, most Single Subject teacher candidates are scored by their student teaching supervisor. Issues with a Supervisor assessing a candidate include: - A Supervisor providing a favorable or unfavorable rating to a teacher candidate based on previous relationship, knowledge, or observation of the teacher candidate. - A Supervisor providing a favorable rating to a teacher candidate because a failing rating would require the teacher candidate (and in turn the Supervisor) to do more work to attempt to remediate the task. However, one might argue that supervisors have more knowledge of a candidate's performance and therefore may provide more accurate appraisals of a candidate's performance. Multiple Subject teacher candidate submissions are scored by assessors other than the candidate's university supervisor because there are enough trained assessors in the multiple subject program to ensure that supervisors do not need to grade their own supervisee's submission.

(continued)

Table 4.6. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
4.15	JS 4.20 Specifications should describe processes for assessing scorer consistency and potential drift over time in raters' scoring.	3	FAST documentation (Response to the Assessment Design Standards, May 2018) does not address processes for assessing scorer drift over time. Processes for monitoring scoring consistency are described in the rationales for 4.9 to 4.11 above.
4.16	JS 4.20 The basis for determining scoring consistency (e.g., percentage of exact agreement, percentage within one score point, or some other index of agreement) are indicated.	5	FAST determines the (a) percentage where scorers disagreed in the overall decision over whether the candidate had passed or failed the project, (b) the percentage of decisions where both scorers assigned the exact same score across tasks, (c) the percentage of decisions where both scorers assigned a score within +1/-1 point across tasks, and (d) the percentage where scorers disagreed over whether the candidate had passed or failed the subtask (described in Appendix G of Response to the Assessment Design Standards, May 2018).
4.17	JS 4.20 The process for selecting, training, qualifying, and monitoring scorers is specified by the [model sponsor].	3	Selecting: FAST uses its teacher faculty to score candidate submissions. There is no formal selection process. Training: The process for training scorers is described in the rationales for 4.1 to 4.6 above. Qualifying: The process for qualifying (calibrating) assessors is described in the rationales for 4.4 to 4.6 above. Monitoring Scorers: The process for monitoring scores is described in 4.9 to 4.11 above. FAST does not currently monitor scorers for scoring consistency *during* the scoring window.
4.18	JS 4.20 To the extent possible, scoring processes and materials anticipate issues that may arise during scoring.	3	Scoring processes and materials do not anticipate some predictable issues that may arise during scoring, such as: - An assessor who cannot successfully calibrate. - An assessor who does not score consistently with the assessor(s) who double score(s) with them and or the third person brought in to adjudicate the scores. In this situation, it is likely the submissions scored by this assessor are not accurate. - A candidate's whose files are unusable (e.g., no audio on a video recording). - A candidate who submits unoriginal work. If there are procedures in place to handle such issues, then those procedures should be included in the model sponsor's scoring documentation.
4.19	JS 6.9 [The model sponsor has] procedures in place to monitor consistency of scoring across administrations (e.g., year-to-year comparability).	NA	FAST just completed its first operational year with revised rubrics. Such a procedure should be specified for the future, but it is not applicable for the 2018–19 academic year.

(continued)

Table 4.6. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
4.20	JS 6.9 [The model sponsor] appropriately retrains, rescores, and dismisses some scorers, and/or reexamines the scoring rubrics or programs based on inaccurate or inconsistent scoring.	3	Retraining occurs during assessor training. If a new assessor does not independently align with ratings made by a group of experts during training, the scorer is paired with a more experienced and calibrated scorer to shadow the scoring process and to discuss the rationale of the experienced scorer in awarding scores, while unofficially scoring the same candidate response (as described in Appendix F, Scorer Training Procedures, of the Response to the Assessment Design Standards, May 2018). Per the guidance in Appendix F of the Response to the Assessment Design Standards, if an experienced assessor does not independently align with ratings made by a group of experts during assessor training, the trainer and assessor discuss the rationale for awarding the score and the trainer clarifies the assessor's misunderstandings. However, the model sponsor indicated that the assessors who participated in the 2017–18 calibration sessions were not required to participate in the 2018–19 calibration sessions [interview with model sponsor on April 27, 2019]. The Response to the Assessment Design Standards document states that "any task that receives a non-passing score on any rubric section is re-scored by a second trained scorer. If there is a discrepancy between the two scores, a third scorer will be used to determine the score (p. 20)." Calibration of scorers is not monitored during the scoring window. Double scoring is analyzed *after* the close of the scoring window and is used for documentation of reliability only. This process does not allow for retraining low reliability scorers *during* the scoring window or dismissal of low reliability scorers. Thus, corrective and preventative actions can only be applied to future scoring windows.
4.21	JS 6.9 Analyses monitor possible effects on scoring accuracy of variables such as scorer, task, time or day of scoring, scoring trainer, scorer pairing, and so on, to inform appropriate corrective or preventative actions.	2	FAST conducts analyses to monitor scoring accuracy; however, FAST does not analyze possible effects on scoring accuracy of variables such as time of day or day of scoring, scoring trainer, scorer pairing, or other more specific variables, to inform appropriate corrective or preventative actions. Double scoring is analyzed *after* the close of the scoring window and is used for documentation of reliability only. This process does not allow for providing feedback to and remediating low reliability scorers *during* the scoring window. Thus, corrective and preventative actions can only be applied to future scoring windows.

(continued)

Table 4.6. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
4.22	JS 6.9 Consistency in applying scoring criteria is checked by independently rescoring randomly selected test responses.	4	Approximately 15% of all submissions are double scored. For the Multiple Subject credential, the submissions to be rescored are randomly selected. Because the Single Subject areas have few assessors, the Assessment Coordinator determines which submissions are double scored (and thus is not random). Across all subject areas, 15% of submissions are double scored. Additional detail on the process for monitoring scoring consistency is provided in the rationales for 4.9 to 4.11.
4.23	JS 6.9 Periodic checks of the statistical properties (e.g., means, standard deviations, percentage of agreement with scores previously determined to be accurate) of scores assigned by individual scorers during a scoring session are used to provide feedback for the scorers, helping them to maintain scoring standards.	1	FAST does not use periodic checks of the statistical properties of scores assigned by individual scorers during a scoring session to provide feedback to the scorers during the scoring window. Due to infrastructure constraints, such a process is difficult. Including at least a mid-scoring window calibration exercise could help identify scorers who are misinterpreting rubrics or not identifying proper evidence when scoring.
4.24	JS 6.9 Those responsible for scoring document the procedures followed for scoring, procedures followed for quality assurance of that scoring, the results of the quality assurance, and any unusual circumstances.	4	FAST documents procedures for scorer training and results of data analyses conducted on double scored submissions (see FAST Response to the Assessment Design Standards, May 2018). Unusual circumstances were not reported.

Note. NA = Not applicable.

Table 4.7. Claim 7 Ratings on the Assessment Design and Joint Standards for FAST

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
7.1	ADS 2(g) The model sponsor conducting scoring for the program provides results on the TPA to the individual candidate based on performance relative to TPE domains and/or to the specific scoring rubrics within a maximum of three weeks following candidate submission of completed TPA responses.	5	Candidates receive results on the TPA relative to the scoring rubrics within the timeline established by the Commission. Scoring rubrics are linked to the TPE elements in the FAST Manual and on the TSP Score Report.
7.2	ADS 2(g) The model sponsor follows the timelines established with programs using a local scoring option for providing scoring results.	5	FAST is a local model itself. Documentation states that FAST provides scores to candidates within 3 weeks (FAST Response to the Assessment Design Standards, May 2018).
7.3	ADS 2(g) The model sponsor provides results to programs based on both individual and aggregated data relating to candidate performance relative to the rubrics and/or domains of the TPEs.	5	FAST is a local model itself so scores are available to them by default. Scores are provided to Multiple and Single Subject Program coordinators for review by their program faculty (FAST Response to the Assessment Design Standards, May 2018). Scoring rubrics are linked to the TPE elements in the FAST Manual and on the TSP Score Report.
7.4	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	The FAST Manual (v2.0) begins with a letter to teacher candidates from the FAST Coordinator. This letter explains that FAST was developed to evaluate mastery of the TPEs and that mastery of TPEs is required for candidates seeking recommendation for a Preliminary California Credential. The letter also states that FAST is just one of the requirements for earning a Preliminary Credential. The letter then describes the FAST scoring rubric, explaining that candidates must obtain a minimum score of '2' (Meets Expectations) on every rubric in order to pass FAST. The letter goes on to explain that candidates can revise and resubmit any section on which they receive a non-passing score. Finally, the letter states that, "A history of your scores will be available to you through Tk20 for sharing with your professional induction program supervisor as you see fit." Furthermore, on pg. 41 of the FAST model there is a formal "Intended Use Policy." The Intended Use Policy states that FAST (a) provides evidence on the pedagogical competence (defined by the TPEs) of Multiple and Single Subject Credential Candidates at Fresno State and (b) provides useful information for determining program quality and effectiveness.

(continued)

Table 4.7. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
			While score use and interpretation guidance is available in the aforementioned materials, the candidate score reports do not include guidance about how scores should be interpreted and used. However, if candidates receive a non-passing score on any rubric, they are contacted by the FAST coordinator (multiple subject) or university coach (single subject) and notified about which rubrics they did not pass and to contact the coordinator/coach to make an appointment to discuss what they need to do to revise the section that they did not pass. Because FAST is a local assessment there is no program score report that is produced, although it's recommended that the FAST model sponsor compute descriptive statistics on rubrics and tasks (within and across credential areas) to help inform program strengths and weaknesses; this type of analysis is in line with a stated purpose of FAST—i.e., determining program quality and effectiveness.
7.5	JS 1.1 The [model sponsor] specifies in clear language the contexts in which test scores are to be employed.	4	The FAST Response to the Assessment Design Standards notes that the FAST has not been released for use by other institutions and that to maintain its validity, it may only be used as designed (p. 13). (See also rationale for 7.4 above). However, this information is not included in score reports.
7.6	JS 1.2 A summary of the evidence and theory bearing on the intended interpretation is presented for each intended interpretation of test scores for a given use. Evidence may come from studies conducted locally, in the setting where the test is to be used; from specific prior studies; or from comprehensive statistical syntheses of available studies meeting clearly specified study quality criteria. No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test score interpretation for a given use determine the value of a particular kind of evidence.	4	An extensive field test of FAST was conducted in 2017–18. The findings from the field test are presented in the Response to the Assessment Design Standards document. The report does not discuss the theory bearing on the intended interpretation of test scores for each given use. Because 2018–19 is the first operational year of the revised FAST, it is likely too soon to expect the model sponsor to have conducted extensive studies at this point. The model sponsor could use findings from the present comparability study to support intended use interpretations.

(continued)

Table 4.7. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
7.7	JS 2.13 The standard error of measurement, both overall and conditional (if reported), is provided in units of each reported score.	NA	Some small and/or specialized assessments cannot be expected to provide data that typically come from larger, more traditional assessment programs. Related to this point, with respect to reliability the *Joint Standards* state, "… there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment (p. 41)." FAST does not provide standard error of measurement; however, see the rationales for *Joint Standards* 3.8, 4.18, 4.20, and 6.9 for a discussion of how the FAST model addresses scorer training, calibration, and monitoring of scoring accuracy. Moreover, by double scoring all non-passing rubric scores, FAST has a built-in safety net for addressing the classification accuracy of pass/fail decisions.
7.8	JS 3.8 [The model sponsor] collects and reports evidence of the validity of constructed response score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores.	5	After scoring ends, FAST compares ethnicity, self-rated English language fluency, self-reported disability, and gender category subgroup subtest and final score performance after each administration using Mann-Whitney *U* and Kruskal-Wallis *H*. Results can be found in Appendix G of Response to the Assessment Design Standards (May, 2018). Small sample sizes of some subgroups (e.g., Limited Working Proficiency in English, Asian, Black) should be accumulated over time to build a large of enough sample prior to conducing these analyses.
7.9	JS 4.22 [The model sponsor] specifies the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.	5	FAST is a criterion-referenced test. Interpreting FAST scores is based on the rubric used for all tasks. The 4-point scale of the rubric ranges from "Does Not Meet Expectations" to "Exceeds Expectations." The model sponsor conducted a Passing Standard Workshop in which a panel of experts reviewed the rubrics and came to consensus agreement that a Level 2 ratings ("Meets Expectations") must be obtained on every rubric in order to pass FAST.

(continued)

Table 4.7. (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
7.10	JS 5.0 [The model sponsor] documents evidence of fairness, reliability, and validity of test scores for their proposed use.	4	After scoring ends, an interrater reliability analysis is conducted to determine the percentage of exact, adjacent, and non-adjacent agreement on each rubric score and for each pass/fail decision on each task. FAST also compares ethnicity, self-rated English language fluency, self-reported disability, and gender category subgroup subtest and final score performance after each administration using Mann-Whitney U and Kruskal-Wallis H. Small sample sizes of some subgroups (e.g., Limited Working Proficiency in English, Asian, Black) should be accumulated over time to build a large of enough sample prior to conducting these analyses. See also rationales for 4.9 to 4.11. Additional evidence of fairness, reliability, and validity of test scores for their proposed use is not available at this early stage (i.e., first operational year).
7.11	JS 5.0 Test scores are derived in a way that supports the interpretations of test scores for the proposed uses of tests.	4	FAST scoring is clear at the indicator level. Each indicator within a rubric relates to a score point. FAST trainers instruct assessors to evaluate each submission on each indicator of each scoring rubric. However, scores are not reported at the indicator level. They are reported at the overall rubric level. If an assessor determines that various rubric indicators are at different levels, they must determine which level the overall rating is to be made based on their own judgment (i.e., holistically). While the FAST representative has indicated that most indicators are weighted equally (interview on April 27, 2019), a short guideline for determining how to weight the importance of individual indicators for the overall rubric rating should be written and provided at future scorer training sessions (particularly to identify indicators like Reflection and Self-Evaluation). If scorers are inconsistently weighting indicators in assigning rubric scores, this could threaten the reliability of scores.
7.12	JS 6.10 Reports and feedback are designed to support valid interpretations and use and minimize potential negative consequences.	3	FAST score reports—for SVP and TSP—provide scores for each rubric (see Appendices 4.F to 4.I). Rubric scores are accompanied by scorer comments for the TSP but not for the SVP. Score reports do not include a total score or an overall pass/fail determination. To help support valid interpretations and to minimize potential negative consequences, FAST should include guidance in the score reports about how rubric level scores should be used. An overall pass/fail determination is also recommended for inclusion. The model sponsor is also encouraged to include guidance in score reports that (a) FAST scores should be used in conjunction with other measures of performance to determine a candidate's preparedness for beginning teaching and (b) that FAST has not been released for use by other institutions and that to maintain its validity, it may only be used as designed. This language is included in the FAST Manual, but it is not included on score reports.

Note. CR = Cannot rate at this time. NA = Not applicable.

Table 4.8. Claim 8 Ratings on the Assessment Design and Joint Standards for FAST

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
8.1	ADS 1(a) Collectively, the tasks and rubrics in the assessment address key aspects of the six major domains of the TPEs.	5	Per the FAST Tasks Matrix, the sponsor has documented the relationship between the elements (i.e., key aspects) of the six major domains of the TPEs and FAST tasks in the FAST Tasks Matrix. Furthermore, this same information is included in the candidate manual and in the TSP Score Report such that all candidates can see which TPE elements (key aspects) are assessed by each FAST rubric.
8.2	ADS 2(a) In relation to the key aspects of the major domains of the TPEs, the pedagogical assessment tasks, rubrics, and the associated directions to candidates are designed to yield enough valid evidence for an overall judgment of each candidate's pedagogical qualifications for a Preliminary Teaching Credential as one part of the requirements for the credential.	5	The SVP and TSP are complex performance tasks that require candidates for a preliminary Multiple Subject or Single Subject teaching credential to perform tasks and activities aligned with the elements (key aspects) of the six TPE domains. Multiple robust rubrics (each measuring multiple TPE elements) are evaluated to judge the submissions, and candidates are required to provide multiple pieces of evidence for each rubric. The FAST Candidate Manual and TSP Score Report informs candidates of the TPE elements (i.e., key aspects) that are measured by each FAST rubric.
8.3	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	See rationale for 7.4 above in Table 4.7. In addition, the FAST Manual states, "A history of your scores will be available to you through Tk20 for sharing with your professional induction program supervisor as you see fit." This suggests that FAST scores can be useful for informing professional development, although such a use is not explicitly stated, nor is this guidance included in score reports. Furthermore, in an interview with the FAST Coordinator on April 27, 2019, the Coordinator indicated that in her opinion FAST is most valuable as a program assessment, although aside from the interrater reliability analyses and equity analysis included in Appendix G of the FAST Response to the Assessment Design Standards, there is no formal analysis or documentation of how aggregate, program-level FAST data is used diagnostically to identify strengths and weaknesses of the program. In terms of informing candidate readiness, the Coordinator indicated that they downplay the scores and focus a lot more on what candidates learn by going through the process. Moreover, during the interview, the FAST Coordinator noted that she meets with every multiple subject candidate that receives a non-passing score on any rubric (single subject candidates who receive a non-passing score on any rubric meet with a faculty person in their discipline area). These candidates then resubmit that section of the project. In 2018–19, 100% of candidates passed FAST when taking into consideration retakes. This suggests that FAST is used primarily for formative purposes at Fresno State.

(continued

Table 4.8 (Continued)

#	Assessment Design/Joint Standard	FAST Rating	Rationale for FAST Rating
8.4	JS 6.10 [Score report] interpretations describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.	3	FAST score reports—one for TSP and one for SVP—provide scores for each rubric. Rubric scores are accompanied by scorer comments for the TSP but not for the SVP. Also, the TSP score report is accompanied by the rubrics, but not the SVP score report. Score reports for both TSP and SVP do not include a total score or an overall pass/fail determination. The precision/reliability of the scores is not presented. Score reports are not accompanied by guidance on how the scores should be used. Thus, it is recommended that FAST include guidance in the score reports about how rubric level scores should be used. An overall pass/fail determination is also recommended for inclusion. In addition, the model sponsor is encouraged to include guidance in score reports that (a) FAST scores should be used in conjunction with other measures of performance to determine a candidate's preparedness for beginning teaching and (b) that FAST has not been released for use by other institutions and that to maintain its validity, it may only be used as designed. This language is included in the FAST Manual, but it is not included on score reports.
8.5	JS 6.10 Score precision [is] depicted by error bands or likely score ranges, showing the standard error of measurement.	NA	Some small and/or specialized assessments cannot be expected to provide data that typically come from larger, more traditional assessment programs. Related to this point, with respect to reliability the Joint Standards state, "… there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment (p. 41)." FAST does not provide standard error of measurement; however, see the rationales for Joint Standards 3.8, 4.18, 4.20, and 6.9 for a discussion of how the FAST model addresses scorer training, calibration, and monitoring of scoring accuracy. Moreover, by double scoring all non-passing rubric scores, FAST has a built-in safety net for addressing the classification accuracy of pass/fail decisions.
8.6	JS 6.10 The interpretive materials prepared by the [model sponsor] address common misuses or misinterpretations.	4	The FAST Manual includes an "Intended Use Policy." The FAST Response to the Assessment Design Standards notes that the FAST has not been released for use by other institutions and that to maintain its validity, it may only be used as designed (p. 13). However, this information is not included in score reports. FAST representatives tell teacher candidates not to share their scores with other candidates and not to "read into" their scores too much.

edTPA Results

Table 4.9. Claim 3 Ratings on the Assessment Design and Joint Standards for edTPA

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
3.1	ADS 1(a) The assessment [includes] multi-level scoring rubrics that are clearly related to the TPEs that the task measures.	4	Rubrics are categorized as Planning, Instruction, and Assessment; see Appendices 4.J, 4.K, and 4.L, respectively, for an overview of each of these rubrics. edTPA uses five-point rubrics that are related to the TPE domains that the tasks measure, per the mapping of TPEs to rubrics in the edTPA Transition Plan (see pg. 31 and pgs. 115-149). However, this information (i.e., rubric to TPE linkage) is not readily available to programs and candidates. Thus, the relation between the edTPA tasks (and rubrics) and the TPEs is not as clear as it could be. For example, the CalTPA Performance Assessment Guides and the FAST Manual and TSP Score Report demonstrate the linkage between TPEs and tasks (and rubrics) for candidates and programs.
3.2	ADS 1(c) The model sponsor defines scoring rubrics so candidates for credentials can earn acceptable scores on the Teaching Performance Assessment with the use of different content-specific pedagogical practices that support implementation of the TK-12 content standards and curriculum frameworks.	4	edTPA's rubrics are either tailored to each credential content domain or are content neutral. Language within the rubrics is general enough to allow candidates to earn acceptable scores with the use of different subject-specific pedagogical practices and curriculum frameworks. However, neither edTPA Handbooks nor score reports provide the linkage between edTPA tasks and rubrics and the TPEs, which are directly and purposely aligned to the TK-12 content standards and curriculum frameworks. Thus, the connection between edTPA scoring rubrics and TK-12 content standards is not as clear as it could be.
3.3	ADS 1(h) The model sponsor develops scoring rubrics that focus primarily on teaching performance.	5	edTPA's rubrics, which include Planning, Instruction, and Assessment, focus primarily on teaching performance.
3.4	ADS 1(h) The model sponsor develops scoring rubrics 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	In the Transition Plan, it was specified that edTPA scorers must successfully complete training and qualification in the prevention of bias before officially scoring edTPA portfolios. In rubric 6 (Learning Environment) of edTPA's Understanding Rubric Level Progressions (Secondary Mathematics edTPA Fall 2016)), edTPA representatives write that "Scorers are cautioned to avoid bias related to their own culturally constructed meanings of respect." Bias training on factors such as personal attire, appearance, demeanor, speech patterns and accents or any other biases not likely to affect job effectiveness and/or student learning is presented via a 10-slide module. Analyses of scores between primary English speakers and primary other language speakers shows no significant difference in scores. edTPA periodically engages a Bias Committee that reviews handbooks and other materials when issues arise. Locations of various types of bias (e.g., content, language, stereotypes) are identified and corrections are made. edTPA representatives held an initial Bias Committee meeting to review all its candidate facing materials prior to use of the assessment.

(continued)

Table 4.9. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
3.5	JS 4.18 Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.	5	Rating scales are operationalized by benchmark portfolios, which are used in scorer training. Furthermore, benchmark materials are regularly updated in response to changes to the edTPA handbooks, as reported in the "edTPA Training Improvement Timeline Final" document. This document also demonstrates that aspects of scorer training are revisited, updated, and improved over time (e.g., scorer training modules are refreshed; recorded lead trainer walking through the score form for review by trainees; redesigned training and qualification for low incidence content areas). In addition to operationalizing each rubric scale with pre-selected benchmarks, a supplementary document, *Thinking Behind the Rubrics* (TBR), is provided to help scorers make distinctions between score levels (including examples of candidate performance). Evidence of scoring reliability is provided using the Cohen Kappa statistics, which demonstrates by rubric the extent to which scorers use the full range of the given 5-point scale beyond estimates of chance (see example Administrative Reports included in the edTPA Transition Plan).
3.6	JS 4.18 [Scoring criteria are] presented by the [model sponsor] with sufficient detail and clarity to maximize the accuracy of scoring.	5	Each rubric describes both the parameters of what is being evaluated and the level of quality (such as an activity's duration, individualization, integration, and connectedness) at which the candidates may perform on each task. Also, the Understanding Rubric Level Progression documents include detailed scoring criteria and clear instructions. The model sponsor provides definitions and guidelines for making scoring decisions. These documents present the score-level distinctions and other information for each edTPA rubric, including: 1) Elaborated explanations for rubric Guiding Questions 2) Definitions of key terms used in rubrics 3) Primary sources of evidence for each rubric 4) Rubric-specific scoring decision rules
3.7	JS 4.18 [The model sponsor provides] multiple examples of responses at each score level for use in training scorers and monitoring scoring consistency. [These] are typically added to scoring specifications during item development and tryouts.	4	edTPA provides an example of each score level in its training modules for each rubric. In addition, assessors see examples in end of module quizzes, practice sets, and calibration sets. Across all 15 rubrics, multiple examples are seen at each score level, but for each rubric, only one example is provided for some score levels.

Table 4.10. Claim 4 Ratings on the Assessment Design and Joint Standards for edTPA

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
4.1	ADS 1(g) The TPA model sponsor [provides] materials appropriate for use by [assessors] to become familiar with the design of the TPA model, the candidate tasks, the scoring rubrics, [and scoring processes].	5	edTPA representatives provide assessors with materials via Pearson's My Learning Bridge application and ePEN. Modules within the Bridge application include an overview of edTPA, training materials related to every rubric, and the portfolio scoring system. Assessors become familiar with the design of edTPA, the candidate tasks, rubrics, and scoring processes. Examples of different score points are provided. Within the scoring application, ePEN, the candidate tasks are also described. All assessors possess their relevant candidate subject handbook, rubrics, and the *Thinking Behind the Rubrics* document. In addition, scorers scoring California submissions are provided with a supplemental resource called, "Deep understanding of the TPEs."
4.2	ADS 1(h) The model sponsor develops 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	A component of the edTPA assessor training is the Preventing Bias module. This slide deck introduces various protected groups and how personal characteristics of teacher candidates can bias assessors. It is emphasized that candidates should not be rated on factors such as personal attire, appearance, demeanor, speech patterns and accents, quality of writing, strong/weak evidence in one area, and classroom setting. Additionally, assessors are asked to notify an edTPA coordinator if they are familiar with a candidate or have strong preferences or associates related to materials or a candidates' characteristics. In these cases, they will be reassigned to another portfolio.
4.3	ADS 2(c) The assessor training program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and the multi-level scoring rubrics.	4	The edTPA Transition Plan (p. 83) indicates that "Supplemental reference material will be provided for California scorers. The supplemental resource, 'Deep understanding of the TPEs' will be available to scorers of California portfolios." Aside from making this supplemental resource available to scorers scoring California portfolios, there is no documentation that demonstrates convincingly that assessors have gained a deep understanding of the TPEs." There is evidence that the assessor training program demonstrates convincingly that assessors gain a deep understanding of the edTPA tasks and rubrics. For example, after completing all modules in the My Learning Bridge application, assessors are asked to complete an independent scoring activity to practice scoring on their own. After this activity, assessors review a recorded interactive group session prior to scoring their qualification portfolios. On the qualification portfolios, assessors must meet edTPA's standards of reliability on separate portfolios. Scorers must reach at least 46% exact agreement with no more than one non-adjacent score. Supervisors and Trainers must reach at least 53% exact agreement with no more than one non-adjacent score. These qualification portfolios replicate the assessor's job and require a deep understanding of edTPA's pedagogical assessment tasks and its multi-level scoring rubrics.

(continued)

Table 4.10. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
4.4	ADS 2(c) The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy and calibration in relation to the scoring rubrics associated with the task.	5	Scorers complete training curriculum composed of over 20 hours of independent, online training modules, as well as independent scoring and the opportunity to discuss any questions about scoring with a trainer before completing the qualification scoring exercises. After completing the self-paced training modules, scorers will score a practice edTPA. Scorers are encouraged to follow up with Trainers during trainers' scheduled office hours to get feedback on the practice calibration. Then, scorers score two submissions to become qualified to score. After the first of the two, scorers are encouraged to review the official scores and annotations and attend Trainer office hours to answer any questions they may have before attempting the second portfolio. Calibration is done electronically with adherence to minimum calibration thresholds. Scorers must reach at least 46% exact agreement with no more than one non-adjacent score. Supervisors and Trainers must reach at least 53% exact agreement with no more than one non-adjacent score. edTPA Trainers require assessors in low-incidence fields to complete different training modules and meet via webinar. Assessors in very low incidence fields (Classical Languages, Education Technology Specialist, and Literacy Specialist) work with other assessors and a trainer to consensus score. It is the trainer's decision when the assessor is qualified.
4.5	ADS 2(c) The model sponsor uses only assessors who successfully calibrate during the required TPA model assessor training sequence.	5	Only assessors who successfully calibrate on two portfolios can score for edTPA. Low incidence subject areas (Agriculture Education, Business Education, Educational Technology Specialist, Family and Consumer Science, Health Education, Library Specialist, Literacy Specialist, Technology and Engineering Education, and Classical Languages) do not calibrate like other areas. Instead, Trainers discuss several practice portfolios that have been consensus scored via webinar (edTPA Transition Plan).
4.6	ADS 2(c) When new pedagogical tasks and scoring rubrics are incorporated into the assessment, the model sponsor provides additional training to the assessors, as needed.	NA	The revisions to edTPA to address the revised ADS and TPEs did not constitute the need to develop new pedagogical tasks or scoring rubrics.
4.7	ADS 2(e) All approved models must include a local scoring option in which the assessors of candidate responses are program faculty and/or other individuals identified by the program who meet the model sponsor's assessor selection criteria. These local assessors are trained and calibrated by the model sponsor, and whose scoring work is facilitated, and their scoring results are facilitated and reviewed by the model sponsor.	5	edTPA offers a local scoring option that allows teacher preparation programs to officially score portfolios from their own campus. These assessors are held to the same standards in calibration and scoring as centralized scorers. All locally scored portfolios are double scored (using edTPA's national sample of assessors). The number of programs that participate in local scoring varies. Because of the time commitment required to score, at most seven or eight local programs have participated in the local scoring option in recent school years. Often local programs will purposely locally score for just a single year as a professional development opportunity to help faculty familiarize with the assessment.

(continued)

Table 4.10. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
4.8	ADS 2(e) The model sponsor must provide an annual audit process that documents that local scoring outcomes are consistent and reliable within the model for candidates across the range of programs using local scoring and informs the Commission where inconsistencies in local scoring outcomes are identified. If inconsistencies are identified, the sponsor must provide a plan to the Commission for how it will address and resolve the scoring inconsistencies both for the current scoring results and for future scoring of the TPA.	5	edTPA offers a local scoring option that allows teacher preparation programs to officially score portfolios from their own campus. edTPA complies with the ADS by requiring that (a) assessors are held to the same standards in calibration and scoring, (b) all locally scored portfolios are double scored (using edTPA's national sample of assessors), and (c) by reporting inconsistencies to the Commission.
4.9	ADS 2(e) The model sponsor provides a detailed plan for establishing and maintaining scorer accuracy and inter-rater reliability during field testing and operational administration of the assessment.	5	edTPA's 2015 Administration Report (found in the edTPA Transition Plan) details edTPA's plan for establishing and maintaining scorer accuracy. At least 10% of portfolios are randomly back scored. Also, operational portfolios that lie within the double scoring band around California's cut score are double scored. A third score is made by scoring supervisors to adjudicate discrepant scores or issues with the rubrics. edTPA's threshold of acceptability for inter-rater reliability is stated as total agreement (exact/adjacent > 90% and kappa n > .80. In 2014, 93.3% of scores were either adjacent or exact matches (edTPA Transition Plan). edTPA researchers found this degree of consistency across rubrics (with a range of 89.9% to 97.1% (between Rubric 2 and 6). Backreading, double scoring, and use of "benchmark"/validity papers occurs throughout the scoring window.
4.10	ADS 2(e) The scoring process conducted by the model sponsor to assure the reliability and validity of candidate outcomes on the assessment may include, for example, regular auditing, selective back reading, and double scoring of candidate responses near the cut score by the qualified, calibrated scorers trained by the model sponsor.	5	edTPA uses Scoring Supervisors, typically former edTPA high performing scorers who have been promoted, to "backread" a percentage of both random candidate portfolios and portfolios near the cut score. Backreading is rescoring a previously scored portfolio for the purpose of reviewing the original score and providing feedback to the scorer. edTPA also uses calibrated assessors (who must calibrate on two separate portfolios to qualify), and routine monitoring of inter-rater reliability to assure the reliability and validity of candidate outcomes. edTPA representatives require assessors to recalibrate on validation portfolios approximately every 100 days.
4.11	JS 3.0 All [scoring procedures steps] are designed in such a manner as to minimize construct-irrelevant variance.	5	There are no apparent scoring procedures to correct that could further minimize construct-irrelevant variance. Bias awareness is presented in a module during training and instructions for scoring are straightforward and clear. Scorers are monitored throughout the scoring window for accuracy and consistency, and scorers that do not meet performance thresholds are provided feedback to ensure continued scoring accuracy and consistency.

(continued)

Table 4.10. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
4.12	JS 3.4 Test takers receive comparable treatment during the [scoring process]. Those responsible for testing adhere to standardized scoring protocols so that test scores will reflect the construct(s) being assessed and will not be unduly influenced by idiosyncrasies in the testing process.	5	edTPA test takers within subject areas receive comparable treatment during the scoring process. Training is provided online in the same format to all trainers and electronic monitoring ensures each portfolio is scored by a calibrated assessor.
4.13	JS 3.8 Adequate training and calibration of scorers is carried out and monitored throughout the scoring process to support the consistency of scorers' ratings for individuals from relevant subgroups. Where sample sizes permit, the precision and accuracy of scores for relevant subgroups also is calculated.	5	Assessors must complete a 20-hour training that includes completing all online modules in order. After completing an overview of rubrics for Task 1, assessors must complete a scoring activity on their own to qualify to score on calibration sets. Monitoring is conducted via random double scoring of 10% of portfolios and those near the cut score. Assessors are required to recalibrate on validation portfolios approximately every 100 days. Precision and accuracy of scores for relevant subgroups is publicly reported.
4.14	JS 3.8 For human scoring, scoring procedures [are] designed with the intent that the scores reflect the examinee's standing relative to the tested construct(s) and are not influenced by the perceptions and personal predispositions of the scorers.	5	A component of the edTPA assessor training is the Preventing Bias module. This slide deck introduces various protected groups and how personal characteristics of teacher candidates can bias assessors. It is emphasized that candidates should not be rated on factors such as personal attire, appearance, demeanor, speech patterns and accents, quality of writing, strong/weak evidence in one area, and classroom setting. Additionally, assessors are asked to notify an edTPA coordinator if they are familiar with a candidate or have strong preferences or associates related to materials or a candidates' characteristics. In these cases, they will be reassigned to another portfolio.
4.15	JS 4.20 Specifications should describe processes for assessing scorer consistency and potential drift over time in raters' scoring.	5	Pearson's ePEN application helps edTPA scoring leads monitor the interrater reliability of scorers over time. edTPA scoring leads are trained to monitor for scoring drift via automatically produced reports in the system. All new scorers are backread by a scoring supervisor and all are flagged for backreading after submitting their first portfolio. Requalification exercises and validity papers are also used to monitor calibration (edTPA Transition Plan). The Quality Management Plan (QMP) for Scoring documents the processes for assessing scoring consistency.
4.16	JS 4.20 The basis for determining scoring consistency (e.g., percentage of exact agreement, percentage within one score point, or some other index of agreement) are indicated.	5	edTPA's threshold of acceptability for inter-rater reliability is stated as total agreement (exact/adjacent > 90%) and kappa n > .80. In 2014, 93.3% of scores were either adjacent or exact matches (as reported in the edTPA Transition Plan).

(continued)

Table 4.10. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
4.17	JS 4.20 The process for selecting, training, qualifying, and monitoring scorers is specified by the [model sponsor].	5	Selecting: Evaluation Systems, edTPA's scoring partner, lists minimum and preferred eligibility requirements and qualifications to become an edTPA rater on its company's website. Raters must either be a 1) Current or Retired Higher Education Faculty, Field Supervisors, Teacher Preparation Program Administrators and other Higher Education Educators at a state-endorsed Teacher Preparation Programs, or 2) a Current or Retired PK-12 Classroom Teacher, Induction or Peer Assistance Mentor/Coach, National Board Certified Teacher (NBCT), School Principal or Other PK-12 Administrator (e.g. Assistant Principal, Dean of Candidates, etc.). For more information see: http://scoreedtpa.pearson.com/become-an-edtpa-scorer/edtpa-scorer-qualifications.html edTPA recruitment goals target a balance of 50% classroom teachers and 50% teacher educators. The number of California based scorers and the balance of classroom teachers to teacher educators is unknown (edTPA Transition Plan). Training and Qualifying: edTPA assessors are required to participate in scorer training before scoring teacher candidate portfolios. Part of the training includes calibration on two separate subject specific candidate portfolios. Monitoring Scorers: Pearson's ePEN application helps edTPA scoring leads monitor the means, standard deviations, and interrater reliability of scorers over time. edTPA scoring leads are trained to monitor for scoring drift and other issues via automatically produced reports in the system. edTPA representatives require assessors to recalibrate on validation portfolios every 100 days.
4.18	JS 4.20 To the extent possible, scoring processes and materials anticipate issues that may arise during scoring.	5	Scoring processes and materials appear to anticipate and address a comprehensive set of issues that may arise during scoring.

(continued)

Table 4.10. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
4.19	JS 6.9 [The model sponsor has] procedures in place to monitor consistency of scoring across administrations (e.g., year-to-year comparability).	5	edTPA has a strong process in place to monitor consistency in scoring within a given administration year (see ratings and rationales for #4.15, #14.16, and #4.17). Those same processes are used each year, thereby contributing to year-to-year comparability. Their process helps to ensure that scorers are applying the same standards overtime and, therefore, guards against scorer drift and, ultimately, group drift. Based on the guidance of its Technical Advisory Committee (TAC), edTPA purposefully replaces validation portfolios annually to ensure that validation portfolios are novel and distinct and are not influenced by familiarity from use in past training administrations (described in QMP). To help ensure year-to-year scoring consistency, any portfolios that are replaced must represent similar, if not exact score profiles across the rubrics, and demonstrate decision consistency (representing high, medium, and low performance) across old and new benchmarks.[28]
4.20	JS 6.9 [The model sponsor] appropriately retrains, rescores, and dismisses some scorers, and/or reexamines the scoring rubrics or programs based on inaccurate or inconsistent scoring.	5	edTPA's Transition Plan states that edTPA retrains and rescores based on inaccurate or inconsistent scoring as discovered through backreading, double scoring, and validation portfolios. A second validation portfolio will be used if a scorer is does not calibrate on an initial validation portfolio within the scoring window. The dismissal of scorers and reexamination of rubrics because of inaccurate or inconsistent scoring is not mentioned in the Transition plan. As addressed in the QMP, edTPA scorers can be and are dismissed due both to their initial qualification results and their ongoing performance that is monitored throughout active scoring. Scorers are monitored and dismissed based on several potential reasons, including poor interrater reliability, scoring too quickly, scorer drift, scorer leniency or stringency, and periods of inactivity. Scorers may also be dismissed for not meeting standards on the embedded portfolios during scoring. Scorers are asked back for each quarter of the year, and the model sponsor reserves the right to not "rehire" scorers based on their quality metrics.
4.21	JS 6.9 Analyses monitor possible effects on scoring accuracy of variables such as scorer, task, time or day of scoring, scoring trainer, scorer pairing, and so on, to inform appropriate corrective or preventative actions.	5	Pearson's ePEN application helps edTPA scoring leads and scorers themselves monitor the means, standard deviations, and interrater reliability/kappa of scorers over time, by scorer, subject area, rubric, etc. edTPA scoring leads are trained to monitor flags for scoring drift and other issues via automatically produced reports in the system (Portfolio Scoring System Training Module, edTPA Transition Plan).
4.22	JS 6.9 Consistency in applying scoring criteria is checked by independently rescoring randomly selected test responses.	5	At least 10% of portfolios are randomly double scored, scores near the cut score must be double scored, and a third score is made by scoring supervisors to adjudicate discrepant scores or issues with the rubrics.

(continued)

[28] Clarification on maintaining year-to-year consistency provided by model sponsor via feedback on draft report; received on 30 July 2019.

Table 4.10. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
4.23	JS 6.9 Periodic checks of the statistical properties (e.g., means, standard deviations, percentage of agreement with scores previously determined to be accurate) of scores assigned by individual scorers during a scoring session are used to provide feedback for the scorers, helping them to maintain scoring standards.	5	Pearson's ePEN application helps edTPA scoring leads and scorers themselves monitor the means, standard deviations, and interrater reliability/kappa of scorers over time, by scorer, content subject area, rubric, etc. edTPA scoring leads and scorers are trained to monitor flags for scoring drift and other issues via automatically produced reports in the system (Portfolio Scoring System Training Module, edTPA Transition Plan).
4.24	JS 6.9 Those responsible for scoring document the procedures followed for scoring, procedures followed for quality assurance of that scoring, the results of the quality assurance, and any unusual circumstances.	5	edTPA documents its operational and quality assurance procedures and results of data analysis related to scoring in annual administrative reports. See also rationale for 4.23.

Note. NA= Not applicable.

Table 4.11. Claim 7 Ratings on the Assessment Design and Joint Standards for edTPA

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
7.1	ADS 2(g) The model sponsor conducting scoring for the program provides results on the TPA to the individual candidate based on performance relative to TPE domains and/or to the specific scoring rubrics within a maximum of three weeks following candidate submission of completed TPA responses.	4	The edTPA provides teacher candidates with score profiles that include scores on each rubric, which includes the rubric language description of the candidate's performance. The score profile focuses on the score point level as described by the rubric. Candidates and programs are not provided with information on the linkage between rubrics and the TPEs. Candidates receive this information through the edTPA portal. They are alerted that their scores are ready via email. Candidates submit their portfolio based on due dates established by their preparation program. Following the due date, edTPA provides scores within three weeks. If a teacher candidate submits early, scores would not necessarily be available in three weeks. edTPA guarantees scores are returned within three weeks of specific cut-off dates.
7.2	ADS 2(g) The model sponsor follows the timelines established with programs using a local scoring option for providing scoring results.	5	edTPA's Transition Plan states that score reports are delivered to candidates, teacher preparation programs, and the state agencies within a three-week turnaround time between candidate portfolio deadlines and reporting of results.
7.3	ADS 2(g) The model sponsor provides results to programs based on both individual and aggregated data relating to candidate performance relative to the rubrics and/or domains of the TPEs.	4	edTPA score reports are provided to educator preparation programs that the candidates indicate as a score recipient during the registration process. Programs receive the score obtained on each of the edTPA rubrics and overall performance information. The results are provided to educator preparation programs through a data file, and through a secure reporting tool (ResultsAnalyzer®), which can be used to generate custom views and reports. A biannual report is also available to programs that contains descriptive statistics about national, state, and program-specific populations. Appendix 4.M provides the edTPA Institution Report Layout. Candidates and programs are not provided with information on the linkage between rubrics and the TPEs.

(continued)

Table 4.11. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
7.4	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	edTPA provides candidates with a Score Profile Interpretation document (see Appendix 4.N in Volume II of this report), which explains how the candidate's scores on each rubric, total edTPA score, and average rubric score were derived; however, it does not explain how each score should be used. edTPA should consider including guidance in the score reports that rubric level scores are provided for formative purposes and that the edTPA total score should be used in conjunction with other measures of performance to determine a candidate's preparedness for beginning teaching. edTPA also provides programs with access to the web-based ResultsAnalyzer® data reporting tool. Per the edTPA Transition Plan, ResultsAnalyzer® allows programs to generate customized test performance reports to inform state policy, research, and state accountability efforts. The edTPA website states that the edTPA is "intended to be used for teacher licensure and to support state and national program accreditation, and to support program renewal." It also states that it is used for "program completion decisions by institutions." Semi-Annual Summary Reports are made available to programs "to assist them in examining the performance of their candidates as compared to the population of candidates taking edTPA within the associated state and nationally (p.97 of Transition Plan).
7.5	JS 1.1 The [model sponsor] specifies in clear language the contexts in which test scores are to be employed.	5	A caution is included in all edTPA Score Profiles: "This assessment was not designed to compare your performance to that of other candidates. Your score is used to compare your knowledge and skills to that required by various states and institutions in compliance with teacher certification requirements." (See also rationale for 7.4 above)

(continued)

Table 4.11. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
7.6	JS 1.2 A summary of the evidence and theory bearing on the intended interpretation is presented for each intended interpretation of test scores for a given use. Evidence may come from studies conducted locally, in the setting where the test is to be used; from specific prior studies; or from comprehensive statistical syntheses of available studies meeting clearly specified study quality criteria. No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test score interpretation for a given use determine the value of a particular kind of evidence.	5	edTPA has substantial research related to the reliability of its scoring and validity of the interpretations intended to be made by scores based on it. See pages 18-23 of the 2016 edTPA Administrative Report for more information. https://secure.aacte.org/apps/rl/res_get.php?fid=3013&ref=edtpa
7.7	JS 2.13 The standard error of measurement, both overall and conditional (if reported), is provided in units of each reported score.	NA	Some small and/or specialized assessments cannot be expected to provide data that typically come from larger, more traditional assessment programs. Related to this point, with respect to reliability the Joint Standards state, "… there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment (p. 41)." The edTPA model sponsor does not provide conditional standard errors of measurement; however, see the rationales for Joint Standards 3.8, 4.18, 4.20, and 6.9 for a discussion of how edTPA addresses scorer training, calibration, and monitoring of scoring accuracy. Moreover, edTPA double scores all portfolios for which the total score falls within the "double scoring band" (see the edTPA QMP). By double scoring all portfolios within this double score band, edTPA has a built-in safety net for addressing the classification accuracy of pass/fail decisions.
7.8	JS 3.8 [The model sponsor] collects and reports evidence of the validity of constructed response score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores.	5	edTPA has substantial research related to the reliability of its scoring and validity of the interpretations intended to be made by scores based on it. Evidence of validity for relevant subgroups (e.g., gender, teaching context, primary language, and level of education) is reported (see edTPA Transition Plan). Also, see pages 18-23 of the 2016 edTPA Administrative Report for more information. https://secure.aacte.org/apps/rl/res_get.php?fid=3013&ref=edtpa

(continued)

Table 4.11. (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
7.9	JS 4.22 [The model sponsor] specifies the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.	5	edTPA is a criterion-referenced test. Interpreting edTPA scores is based on the rubric used for all tasks. The model sponsor describes the work of a multi-step standard-setting process in summer 2013. See edTPA Transition Plan or this page of edTPA's website for more information: http://www.edtpa.com/PageView.aspx?f=GEN_PerformanceStandard.html
7.10	JS 5.0 [The model sponsor] documents evidence of fairness, reliability, and validity of test scores for their proposed use.	5	edTPA has substantial research related to the reliability of its scoring and validity of the interpretations intended to be made by scores based on it. See pages 18-23 of the 2016 edTPA Administrative Report for more information. https://secure.aacte.org/apps/rl/res_get.php?fid=3013&ref=edtpa The assessment was field tested in 2012 and 2013 (See the 2013 edTPA Field Test: Summary Report). Field test results were used to improve the assessment. Appendix 1 of the edTPA Transition Plan provides the model's description of how it adheres to the Assessment Design Standards.
7.11	JS 5.0 Test scores are derived in a way that supports the interpretations of test scores for the proposed uses of tests.	5	edTPA score reports provide scores for each rubric, total score, and average rubric score (see Appendix 4.N in Volume II of this report). The total score draws from a larger set of performances to create a more reliable measure on which to base pass/fail decisions.
7.12	JS 6.10 Reports and feedback are designed to support valid interpretations and use and minimize potential negative consequences.	4	edTPA score reports provide scores for each rubric, total score, and average rubric score (see Appendix 4.N in Volume II of this report). The total score draws from a larger set of performances to create a more reliable measure on which to base pass/fail decisions. To help support valid interpretations and to minimize potential negative consequences, edTPA should consider including guidance in the score reports that rubric level scores are provided for formative purposes and that the edTPA total score should be used in conjunction with other measures of performance to determine a candidate's preparedness for beginning teaching. Additionally, it would be informative to include the overall pass/fail decision directly on the reports.

Table 4.12. Claim 8 Ratings on the Assessment Design and Joint Standards for edTPA

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
8.1	ADS 1(a) Collectively, the tasks and rubrics in the assessment address key aspects of the six major domains of the TPEs.	4	edTPA tasks and rubrics, revised in 2016, are linked to the six major domains of the TPEs (see Transition Plan, p. 31) and to the TPE elements (see Transition Plan pgs. 115-149). However, this information is not readily available to candidates and programs as it is not included in Handbooks or Score Reports, or other supporting materials. Thus, it is not readily apparent to candidates and programs how edTPA addresses the TPE domains or the key aspects (elements) of the TPE domains.
8.2	ADS 2(a) In relation to the key aspects of the major domains of the TPEs, the pedagogical assessment tasks, rubrics, and the associated directions to candidates are designed to yield enough valid evidence for an overall judgment of each candidate's pedagogical qualifications for a Preliminary Teaching Credential as one part of the requirements for the credential.	4	See rationale for 8.1 above.
8.3	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	See rationale for 7.4 above in Table 4.11, which states that edTPA should consider including guidance in the score reports that rubric level scores are provided for formative purposes and that the edTPA total score should be used in conjunction with other measures of performance to determine a candidate's preparedness for beginning teaching. See also rationale below for 8.4. In addition, edTPA Score Profiles include a "Performance Description" section in which candidates are provided feedback on their performance on each rubric, which can be used diagnostically for identifying strengths and weaknesses. Furthermore, edTPA provides handbook-specific documents, *Understanding Rubric Level Progressions,* which further supports understanding and use of rubrics and rubric scores. Moreover, edTPA has a candidate and faculty version of *Review and Guidance for Low-Scoring Candidates* document that lists common reasons for low scores to assist faculty in counseling candidates who need to retake the assessment.
8.4	JS 6.10 [Score report] interpretations describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.	4	edTPA provides candidates with a Score Profile Interpretation Document (see Appendix 4.N in volume). It includes each rubric score, average rubric score, and total score. It does not include a pass/fail determination. Instead, it includes a link to a website where candidates can look up the passing score for their state. A score use statement is included in all edTPA Score Profiles: "Your edTPA Score Profile is for your records only. This document may not be used to gain certification. States must receive scores from the Evaluation Systems group of Pearson to fulfill certification

(continued)

Table 4.12 (Continued)

#	Assessment Design/Joint Standard	edTPA Rating	Rationale for edTPA Rating
8.4	(continued)		requirements. This assessment was not designed to compare your performance to that of other candidates. Your score is used to compare your knowledge and skills to that required by various states and institutions in compliance with teacher certification requirements." Performance descriptions accompany each rubric score. edTPA should consider including additional guidance in its score reports that rubric scores are provided for formative purposes to help identify strengths and weaknesses, but that the overall total score should be used, in conjunction with other measures, to determine a candidate's preparedness for beginning teaching. Precision/reliability of the scores is not provided in score reports. There is currently no separate sample score report contextualized for California.
8.5	JS 6.10 Score precision [is] depicted by error bands or likely score ranges, showing the standard error of measurement.	NA	Some small and/or specialized assessments cannot be expected to provide data that typically come from larger, more traditional assessment programs. Related to this point, with respect to reliability the Joint Standards state, "… there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment (p. 41)." The edTPA model sponsor does not provide conditional standard errors of measurement; however, see the rationales for Joint Standards 3.8, 4.18, 4.20, and 6.9 for a discussion of how edTPA addresses scorer training, calibration, and monitoring of scoring accuracy. Moreover, edTPA double scores all portfolios for which the total score falls within the "double scoring band" (see the edTPA QMP). By double scoring all portfolios within this double score band, edTPA has a built-in safety net for addressing the classification accuracy of pass/fail decisions.
8.6	JS 6.10 The interpretive materials prepared by the [model sponsor] address common misuses or misinterpretations.	5	A caution is included in all edTPA Score Profiles: "Your edTPA Score Profile is for your records only. This document may not be used to gain certification. States must receive scores from the Evaluation Systems group of Pearson to fulfill certification requirements. This assessment was not designed to compare your performance to that of other candidates. Your score is used to compare your knowledge and skills to that required by various states and institutions in compliance with teacher certification requirements. It may be appropriate to provide guidance about if and how scores should be listed on teaching resumes or used elsewhere outside of the degree program by teacher candidates who pass."

CalTPA Results

Table 4.13. Claim 3 Ratings on the Assessment Design and Joint Standards for CalTPA

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
3.1	ADS 1(a) The assessment [includes] multi-level scoring rubrics that are clearly related to the TPEs that the task measures.	5	CalTPA's two instructional cycles are subject-specific and directly address the TPEs: Cycle 1 (Learning About Students and Planning Instruction) and Cycle 2 (Assessment-Driven Instruction). CalTPA lists the specific TPE elements (i.e., key aspects) aligned with each cycle in its five-point analytic rubrics for both Cycles 1 and 2. The linkage between Cycles and rubrics and TPE elements is included in the Performance Assessment Guides. Following field test (2017–18), CalTPA developers deleted three rubrics and made changes to all CalTPA essential questions and rubrics. Wording changes improved the clarity and language consistency of the rubrics. While most of the rubric language on the deleted rubrics was moved to other rubrics, a few major changes occurred in the process. Incorporation of technology was emphasized in Cycle 2 by making a dedicated technology rubric. Other changes were made to (a) better distinguish score points at the 2, 3, and 4 levels, (b) link score levels with the candidate guidebook, (c) and describe task expectations (e.g., page limits were made). CalTPA designers also enhanced the glossary in the model's assessment materials to help with terminology related to the rubrics. For the 2019–20 year, CalTPA designers plan to make only small wording changes to the model's rubrics.
3.2	ADS 1© The model sponsor defines scoring rubrics so candidates for credentials can earn acceptable scores on the Teaching Performance Assessment with the use of different content-specific pedagogical practices that support implementation of the TK-12 content standards and curriculum frameworks.	5	There are separate guides (including rubrics) for Multiple Subject and Single Subject. Within those, the CalTPA rubrics are neutral with regard to subject-specific pedagogical practices. Language within the rubrics is general enough to allow candidates to earn acceptable scores with the use of different subject-specific pedagogical practices and curriculum frameworks. The CalTPA Performance Assessment Guides show how the CalTPA rubrics are mapped to the TPE elements, which are directly and purposely aligned to the TK-12 content standards and curriculum frameworks.
3.3	ADS 1(h) The model sponsor develops scoring rubrics that focus primarily on teaching performance.	5	Both Cycle 1 and 2 rubrics are aligned with the TPE elements, which are primarily focused on teaching performance (see Appendix 4.O and 4.P for Cycle 1 and Cycle 2 rubrics, respectively, for the Multiple Subject credential).

(continued)

Table 4.13. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
3.4	ADS 1(h) The model sponsor develops scoring rubrics 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	CalTPA's scoring rubrics only relate to pedagogical competence. Other factors such as personal attire, appearance, demeanor, speech patterns and accents are not evaluated. A bias prevention presentation and discussion is included in training. Presenters listed diverse types of bias that may impact scoring and invited attendees to discuss bias and its possible impact in pairs. In August 2018, CalTPA created a four-person bias review committee to ensure draft performance assessment materials were free from potential bias. Members reviewed the CalTPA Performance Assessment Guides, the CalTPA Glossary, identified potential sources of bias, and recommended revisions. Criteria for the review included content, language, offense, stereotypes, fairness, and diversity. While not a candidate factor, differences in quality of teacher candidate placement was perceived as a fairness issue by CalTPA Field Test Coordinators (based on interviews and focus groups conducted). The concern was that not all focus student types were available in some placements. In suboptimal placements, flexibility was desired in choosing Focus student 3. There was also concern about differing educational technology levels in each school, which is a dimension scored in Cycle 2.
3.5	JS 4.18 Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.	5	Beginning with assessor training sessions in the fall of 2018, Trainers distributed a document called "Scorer Process Flow" to help guide rubric decisions (see Appendix 4.Q for an example). Included in the document is a flow chart providing step-by-step instructions for awarding scores to candidates. Assessors and trainers in observed sessions were generally very positive that these charts were a helpful scoring aid. Before the "Scorer Process Flow" document, sample assessor notes and "look fors" were provided to assessors to help them rate during the 2017–18 field test year. "Look fors" included guiding questions (i.e., Where does the evidence come from? What is the evidence?). The "Scorer Process Flow" improves the materials provided to assessors by removing differences in annotations made by lead assessors thus making them consistent across credential areas. In-person assessor training offers assessor trainees an opportunity to develop scores, compare their rationale with peers, and receive specific feedback from Trainers. Assessors must be attuned to the nuances of each rubric to properly score submissions. Most rubrics' level 1 descriptions use conjunctions, which require some extra attention to ensure no credit should be awarded. At levels 4 and 5, assessors must know the requirements of the next lower level to determine if all requirements were met in addition to the requirements at levels 4 or 5. There is no scaling involved with scoring. All rubrics are on a 5-point scale.

(continued)

Table 4.13. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
3.6	JS 4.18 [Scoring criteria are] presented by the [model sponsor] with sufficient detail and clarity to maximize the accuracy of scoring.	5	The rubrics describe both the parameters of what is being evaluated and the level of quality (such as an activity's duration, individualization, completeness, integration, and connectedness) at which the candidates may perform on each task. The rubrics were revised for the 2018–19 school year (operational) to address points of ambiguity on the field test rubrics used in 2017–18 [interview with model sponsor, December 14, 2018].
3.7	JS 4.18 [The model sponsor provides] multiple examples of responses at each score level for use in training scorers and monitoring scoring consistence. [These] are typically added to scoring specifications during item development and tryouts.	4	CalTPA created a standardized sequence of marker papers for assessor trainees to learn how to score. Every assessor trainee reviewed a mid-range submission, a high range submission, a very low range submission, and finally another two mid-range submissions. Participants begin by being told the scores, then use a consensus model where they talk through the scores as a group. The sixth submission is a calibration. Consensus sets are used in small, facilitated groups to demonstrate the process of determining scores, compare scores and rationales with peers, and receive Trainer feedback. Calibration sets are individually scored by assessor trainees to demonstrate their understanding of the scoring rules and qualify for operational scoring. Across all rubrics, multiple examples of responses are presented at each score level for levels 2, 3, and 4, but for each rubric, only one example is provided for some score levels. Examples of performances at the 1 and 5 level are rare for this TPA model.

Table 4.14. Claim 4 Ratings on the Assessment Design and Joint Standards for CalTPA

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
4.1	ADS 1(g) The TPA model sponsor [provides] materials appropriate for use by [assessors] to become familiar with the design of the TPA model, the candidate tasks, the scoring rubrics, [and scoring processes].	5	CalTPA Trainers provide assessors with presentations and materials to review in advance of training and during the training related to the design of the TPA model, the candidate tasks, the scoring rubrics, and scoring processes. These materials familiarize assessors with CalTPA's Instructional Cycle 1 and 2 structure, focal students, subject-specific pedagogy, and philosophy. With this information, assessors become familiar with the design of CalTPA, the candidate tasks, rubric format and content, and scoring processes.
4.2	ADS 1(h) The model sponsor develops 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	CalTPA assessor training includes an activity where assessors discuss bias and its possible impact on assessment scores and several PowerPoint slides on the topic of bias. This in-person training emphasizes that candidates should not be rated on factors such as gender, ethnicity, race, or writing ability. Assessors are instructed to score based on evidence—not candidate characteristics, response characteristics, or scoring conditions. Rubrics focus on teaching-related performance. In 2018–19, the scoring program used by CalTPA, ePEN, was configured so that assessors had to specify evidence from a submission (i.e., tag evidence) to justify scoring. If an assessor doesn't specify evidence, they can't continue to score other rubrics. In 2017–18, tagging was a function that could be used but was not a requirement. When differences in scores occur during live scoring, CalTPA lead assessors examine the tagged evidence to determine where the issue is and to provide feedback to scorers.
4.3	ADS 2(c) The assessor training program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and the multi-level scoring rubrics.	5	The CalTPA rubrics are mapped to the TPE elements. So, as assessors are gaining deep understanding of the rubrics, they're also increasing their understanding of the TPEs. During the 2018–19 operational scoring year, assessors were also asked to score a calibration portfolio using ePEN during training. If they passed, they were cleared to score. If not, they could try to calibrate on a second calibration submission. All calibration was done on-site (with this iteration only allowing two chances to calibrate—rather than 3). Calibration criteria rigor was similar to the field test. For Cycle 1, which has 8 rubrics, trainees could (a) only score non-adjacent on one rubric and (b) had to score with exact agreement on 3 rubrics. For Cycle 2, which has 9 rubrics, trainees could (a) only score non-adjacent on one rubric and (b) had to score with

(continued)

Table 4.14. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
4.3	(continued)		exact agreement on 4 rubrics (CalTPA Scoring QMP). When trainees qualified on the first calibration set, they didn't need to score the second. CalTPA scripted out what lead assessors should say and do to remediate assessors who did not calibrate on the first submission to improve calibration performance. In 2018–19, assessor calibration was high. Roughly 95% of assessor trainees calibrated at training, which was higher than in 2017–18 (interview with Amy Reising December 14, 2018). HumRRO independently analyzed survey data collected from CalTPA trainees (n = 179) who took an assessor post-training survey. Overall, assessors' perceptions of the assessor training were decidedly positive. Specifically, 85 percent of respondents indicated that the CalTPA Program Overview Webinar was helpful (n = 146), 89 percent of respondents indicated that the Prevention of Bias in CalTPA Orientation Webinar was helpful (n = 159), and 87 percent of respondents indicated that the ePEN informational module and videos were helpful (n = 154). Additionally, 97 percent of respondents felt supported during the assessor training process (n = 172) and 98 percent reported that the assessor training they received adequately prepared them for the task of assessing candidates' submissions (n = 175).
4.4	ADS 2(c) The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy and calibration in relation to the scoring rubrics associated with the task.	5	See rationale for 4.3 above.
4.5	ADS 2(c) The model sponsor uses only assessors who successfully calibrate during the required TPA model assessor training sequence.	5	Only assessors who successfully calibrate can score for the CalTPA. Calibration requirements are clearly communicated to the assessors undergoing training. For Cycle 1, which has 8 rubrics, trainees 1) could only score non-adjacent on one rubric and 2) had to score with exact agreement on 3 rubrics. For Cycle 2, which has 9 rubrics, trainees 1) could only score non-adjacent on one rubric and 2) had to score with exact agreement on 4 rubrics (CalTPA Scoring QMP).
4.6	ADS 2(c) When new pedagogical tasks and scoring rubrics are incorporated into the assessment, the model sponsor provides additional training to the assessors, as needed.	5	After the field test, CalTPA developers deleted three rubrics from Cycles 1 and 2 and made changes to all CalTPA essential questions and rubrics. Wording changes improved the clarity and language consistency of the rubrics. While most the rubric language on the deleted rubrics was moved to other rubrics, a few major changes occurred in the process. All assessors, including returning assessors, were retrained/trained with a full, updated training sessions in 2018–19.

(continued)

Table 4.14. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
4.7	ADS 2(e) All approved models must include a local scoring option in which the assessors of candidate responses are program faculty and/or other individuals identified by the program who meet the model sponsor's assessor selection criteria. These local assessors are trained and calibrated by the model sponsor, and whose scoring work is facilitated, and their scoring results are facilitated and reviewed by the model sponsor.	CR	CalTPA will allow faculty at a program to officially score portfolios from their own campus using its centralized scoring platform (Assessor Orientation) beginning in July 2019. Assessors will be trained using the same methods and held to the same standards in calibration and scoring as centralized scoring.
4.8	ADS 2(e) The model sponsor must provide an annual audit process that documents that local scoring outcomes are consistent and reliable within the model for candidates across the range of programs using local scoring and informs the Commission where inconsistencies in local scoring outcomes are identified. If inconsistencies are identified, the sponsor must provide a plan to the CTC for how it will address and resolve the scoring inconsistencies both for the current scoring results and for future scoring of the TPA.	CR	CalTPA will allow faculty at a program to officially score portfolios from their own campus using its centralized scoring platform (Assessor Orientation) beginning in July 2019. Assessors will be trained using the same methods and held to the same standards in calibration and scoring as centralized scoring.
4.9	ADS 2(e) The model sponsor provides a detailed plan for establishing and maintaining scorer accuracy and inter-rater reliability during field testing and operational administration of the assessment.	5	CalTPA has established a Scoring Quality Management Plan document. It states that CalTPA will use only calibrated assessors (who must qualify by meeting a threshold of inter-rater reliability). It also uses scoring experts to backread a percentage of candidate portfolios to check exact and adjacent percentage agreement, scoring rate, and scoring quantity to assure the reliability and validity of candidate outcomes. CalTPA's threshold of acceptability for inter-rater reliability of operational scoring during monitoring is 50% exact agreement with one or less non-adjacent score per 8 rubrics for Cycle 1 and 44.4% exact agreement with one or less non-adjacent score per 9 rubrics (CalTPA Scoring QMP). During operational scoring, CalTPA collects and monitors double scoring inter-rater reliability; assessors who do not meet the quality monitoring standard for inter-rater reliability after 5 double-scored submissions are flagged for back-reading. Additionally, assessors are considered "calibrated" for only a set period of time if they are inactive. After extended time off, they must review scoring documents and recalibrate on a consensus submission (CalTPA Scoring QMP).

(continued)

Table 4.14. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
4.10	ADS 2(e) The scoring process conducted by the model sponsor to assure the reliability and validity of candidate outcomes on the assessment may include, for example, regular auditing, selective back reading, and double scoring of candidate responses near the cut score by the qualified, calibrated scorers trained by the model sponsor.	5	CalTPA uses calibrated assessors (who must qualify by meeting a threshold of inter-rater reliability). It also requires assessors to randomly double score 10% of candidate submissions for the purpose of inter-rater reliability, requires validity (reliability) scoring (i.e., pre-scored submissions sent to all assessors to check the calibration of assessors), and monitors scoring rate and scoring quantity. Note: All submissions that don't meet the passing threshold are read by another assessor and adjudicated by a lead assessor if there is a lack of agreement between the scores assigned by the assessors. See also the rationale for 4.9.
4.11	JS 3.0 All [scoring procedures steps] are designed in such a manner as to minimize construct-irrelevant variance.	5	Scoring rubrics and the "Scorer Process Flow" instructions are detailed and clearly identify the information that should be considered when determining scores. Example responses further solidify this focus on relevant characteristics of a candidate portfolio. In addition, bias training and detailed feedback during training highlight potential sources of construct-irrelevant bias such as the appearance or speech pattern of a candidate. Scorers are monitored throughout the scoring window for accuracy and consistency, and scorers that do not meet performance thresholds are provided feedback to ensure continued scoring accuracy and consistency. There are no apparent scoring procedures to correct that could further minimize construct-irrelevant variance.
4.12	JS 3.4 Test takers receive comparable treatment during the [scoring process]. Those responsible for testing adhere to standardized scoring protocols so that test scores will reflect the construct(s) being assessed and will not be unduly influenced by idiosyncrasies in the testing process.	5	CalTPA test takers receive as comparable treatment during the scoring process as possible. Standardized scoring protocols are standardized, and assessors are trained to disregard construct-irrelevant variance, such as the appearance of a candidate or if a candidate's approach differs from the teaching approach the assessor might have preferred. While the content of a candidate's portfolio might be impacted by the environment in which that candidate prepares a portfolio (e.g., availability of classroom technology), the scoring rubrics and training provide specific guidance to apply scoring rules consistently. Training is provided in person in the same format to all trainers and electronic monitoring ensures each portfolio is scored by a calibrated assessor.

(continued)

Table 4.14. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
4.13	JS 3.8 Adequate training and calibration of scorers is carried out and monitored throughout the scoring process to support the consistency of scorers' ratings for individuals from relevant subgroups. Where sample sizes permit, the precision and accuracy of scores for relevant subgroups also is calculated.	5	Assessors must complete a 2-day training. The training begins with a whole group session to (a) introduce expectations and goals for the training, (b) explain CalTPA's structure, (c) discuss and raise awareness of the possibility of bias in scoring, and (d) assist attendees with ePEN access. Following the whole group session, attendees break into small groups for a credential and cycle specific training and must pass a calibration before training ends to score. Monitoring is conducted via backreading and double scoring of submissions. CalTPA representatives calculate demographic and total score descriptive performance statistics (number, percent, mean, standard deviation, and median, minimum, maximum) by gender, ethnicity, language (e.g., English Only or multilingual), and setting (i.e., urbanicity) subgroups.
4.14	JS 3.8 For human scoring, scoring procedures [are] designed with the intent that the scores reflect the examinee's standing relative to the tested construct(s) and are not influenced by the perceptions and personal predispositions of the scorers.	5	CalTPA assessor training includes a bias discussion activity and several PowerPoint slides related to bias. It is emphasized that candidates should not be rated on factors such as gender, ethnicity, race, or writing ability. Assessors are instructed to score based on evidence—not candidate characteristics, response characteristics, or scoring conditions. Rubrics focus on teaching-related performance. In August 2018, CalTPA created a four-person bias review committee to ensure draft performance assessment materials were free from potential bias. Members reviewed the CalTPA Performance Assessment Guides, the CalTPA Glossary, identified potential sources of bias, and recommended revisions. Criteria for the review included content, language, offense, stereotypes, fairness, and diversity.
4.15	JS 4.20 Specifications should describe processes for assessing scorer consistency and potential drift over time in raters' scoring.	5	CalTPA has established a Scoring Quality Management Plan to ensure scoring consistency. In conjunction with Pearson's ePEN application, CalTPA scoring leads monitor the interrater reliability of scorers over time. Scoring leads are trained to monitor for scoring drift via automatically produced reports in the system. Additionally, assessors are considered "calibrated" for only a set period of time if they are inactive. After extended time off, they must review scoring documents and recalibrate on a consensus submission (CalTPA Scoring QMP).

(continued)

Table 4.14. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
4.16	JS 4.20 The basis for determining scoring consistency (e.g., percentage of exact agreement, percentage within one score point, or some other index of agreement) are indicated.	5	CalTPA's threshold of acceptability for inter-rater reliability for calibration is slightly different between Cycle 1 and Cycle 2. For Cycle 1, which has 8 rubrics, trainees could (a) only score non-adjacent on one rubric and (b) had to score with exact agreement on 3 rubrics. For Cycle 2, which has 9 rubrics, trainees could (a) only score non-adjacent on one rubric and (b) had to score with exact agreement on 4 rubrics (CalTPA Scoring QMP). During operational scoring, CalTPA collects and monitors double scoring inter-rater reliability; assessors who do not meet the quality monitoring standard for inter-rater reliability after 5 double-scored submissions are flagged for back-reading.
4.17	JS 4.20 The process for selecting, training, qualifying, and monitoring scorers is specified by the [model sponsor].	5	Selecting: CalTPA selection requirements include experience teaching in a specific subject area or University experience teaching or supervising TK-12 classroom teachers. In 2018–19, 50% of the 400 trained and calibrated assessors were faculty from institutions of higher education and 50% were TK-12 teachers (May 2019 Standard-Setting Meeting). For more information see: http://www.ctcexams.nesinc.com/TestView.aspx?f=CACBT_Scoring_CalTPA.html Training: See rationales for 3.5, 3.7, 4.1, 4.2, 4.3, 4.6, and 4.13. Qualifying: See rationales for 4.3, 4.5 and 4.13. Monitoring scorers: See rationales for 4.9, 4.10, 4.13, 4.15, 4.16, 4.20 to 4.24.
4.18	JS 4.20 To the extent possible, scoring processes and materials anticipate issues that may arise during scoring.	5	Scoring processes and materials appear to anticipate and address a comprehensive set of issues that may arise during scoring.
4.19	JS 6.9 [The model sponsor has] procedures in place to monitor consistency of scoring across administrations (e.g., year-to-year comparability).	NA	CalTPA just completed its first operational year with revised rubrics. Such a procedure should be specified for the future, but it is not applicable for the 2018–19 academic year.
4.20	JS 6.9 [The model sponsor] appropriately retrains, rescores, and dismisses some scorers, and/or reexamines the scoring rubrics or programs based on inaccurate or inconsistent scoring	5	Documentation indicates CalTPA retrains, rescores, and dismisses assessors, and/or reexamines the scoring rubrics based on inaccurate or inconsistent scoring. Assessors that need coaching can be locked out of the system until they are coached to fix the issue(s). When assessors can't remediate, they are locked out of the scoring system and their contract is not renewed.

(continued)

Table 4.14. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
4.21	JS 6.9 Analyses monitor possible effects on scoring accuracy of variables such as scorer, task, time or day of scoring, scoring trainer, scorer pairing, and so on, to inform appropriate corrective or preventative actions.	5	Pearson's ePEN application helps CalTPA scoring leads monitor the means, standard deviations, and interrater reliability of assessors over time, by assessor, subject area, rubric, etc. Scoring leads members are trained to monitor for scoring drift and other issues via automatically produced reports in the system. During operational scoring, CalTPA collects and monitors double scoring inter-rater reliability and validity (reliability) scoring.
4.22	JS 6.9 Consistency in applying scoring criteria is checked by independently rescoring randomly selected test responses.	5	Pearson's ePEN application provides a mechanism to randomly allow assessors to independently rescore (aka, double score) a submission for the intent purpose of monitoring scoring consistency. A minimum of 10% of submissions are randomly double scored (CalTPA Scoring QMP).
4.23	JS 6.9 Periodic checks of the statistical properties (e.g., means, standard deviations, percentage of agreement with scores previously determined to be accurate) of scores assigned by individual scorers during a scoring session are used to provide feedback for the scorers, helping them to maintain scoring standards.	5	Pearson's ePEN application helps CalTPA scoring leads monitor the means, standard deviations, and interrater reliability of assessors during a scoring session. Scoring leads are trained to monitor for scoring drift and other issues via automatically produced reports in the system. Also, in 2018–19, the ePEN application was configured so that assessors had to specify evidence from a submission (i.e., tag evidence) to justify scoring. If an assessor doesn't specify evidence, they can't continue to score other rubrics. When differences in scores occur during live scoring, CalTPA lead assessors examine the tagged evidence to determine where the issue is and to provide feedback to scorers.
4.24	JS 6.9 Those responsible for scoring document the procedures followed for scoring, procedures followed for quality assurance of that scoring, the results of the quality assurance, and any unusual circumstances.	5	CalTPA has established a Scoring Quality Management Plan to ensure scoring consistency and provides the Commission with a monthly report that includes the model's number of (a) assessors, (b) scored submissions, (c) double scored submissions, (d) pass/fail submissions, (e) types of condition codes, (f) backreads, and (g) backread conferences. The CalTPA model plans to provide formal documentation when the Commission requests annual TPA model administrative reports. See also rationale for 4.23.

Note. CR= Cannot rate at this time.; NA= Not applicable.

Table 4.15. Claim 7 Ratings on the Assessment Design and Joint Standards for CalTPA

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
7.1	ADS 2(g) The model sponsor conducting scoring for the program provides results on the TPA to the individual candidate based on performance relative to TPE domains and/or to the specific scoring rubrics within a maximum of three weeks following candidate submission of completed TPA responses.	5	In 2018–19, CalTPA rubric-level and total score results were reported to candidates directly by the CalTPA sponsor within three weeks. Rubric-level results are mapped to TPE elements, allowing candidates to readily determine their strengths and weaknesses relative to the TPE elements.
7.2	ADS 2(g) The model sponsor follows the timelines established with programs using a local scoring option for providing scoring results.	CR	Local scoring not yet implemented in 2018–19.
7.3	ADS 2(g) The model sponsor provides results to programs based on both individual and aggregated data relating to candidate performance relative to the rubrics and/or domains of the TPEs.	5	CalTPA provides programs with candidate information in a file that includes fields such as CalTPA subject and cycle, rubric scores, total score for cycle, average rubric score obtained for all scored rubrics, and California pass/fail status. Rubric scores are mapped to TPE elements. Analysis of candidate performance across candidate- and aggregate-level data with state averages are provided to programs through a propriety program, ResultsAnalyzer®. The revision of CalTPA was spurred in part because the model's new analytic rubrics provide detailed feedback to allow targeted coaching for teacher candidates and help programs determine how to better align their instruction and curriculum to the TPEs.
7.4	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	The "Introduction to CalTPA's Performance Assessment Guides" clearly states how test scores are intended to be interpreted and used. The CalTPA Performance Assessment Overview (Version 02) document states that the CalTPA is one of multiple measures to inform candidate preparedness. It goes on to state that CalTPA is intended to provide both a formal assessment of candidate ability and a framework of performance-based guidance to inform candidate preparation and continued professional growth. Furthermore, it states that feedback provided at the completion of each cycle is intended to facilitate preparation for the subsequent assessment cycle and that data is shared with institutions to assist them in making program improvements and to guide induction programs as they work with new teachers to individualize learning plans. Score reports include a section called "Understanding Your CalTPA Assessment Results Report," which is included in this report in Appendix 4.R. However, the interpretation and score use guidance that's discussed

(continued)

Table 4.15. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
7.4	(continued)		in the aforementioned materials is not included here, although under "Rubric Performance Summary" it does state that "this information may help you identify your relative strengths and areas for improvement." In addition to including the interpretation and score use information noted above, we also recommend that CalTPA consider including guidance in the score reports that rubric level scores are provided for formative purposes and that the CalTPA total score should be used in conjunction with other measures of performance to determine a candidate's preparedness for beginning teaching. See also use of ResultsAnalyzer® described above in the rationale for 7.3.
7.5	The [model sponsor] specifies in clear language the contexts in which test scores are to be employed.	5	The Candidate Score Report states that the Results Report "is for your records only" and that, "This assessment was not designed to compare your performance to that of other candidates. Your score is used to compare your performance to the performance level set by the Commission on Teacher Credentialing." (See also rationale for 7.4 above)
7.6	JS 1.2 A summary of the evidence and theory bearing on the intended interpretation is presented for each intended interpretation of test scores for a given use. Evidence may come from studies conducted locally, in the setting where the test is to be used; from specific prior studies; or from comprehensive statistical syntheses of available studies meeting clearly specified study quality criteria. No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test score interpretation for a given use determine the value of a particular kind of evidence.	5	An extensive pilot test (2017) and field test (2017–18) of CalTPA was conducted. Findings from the pilot test are presented in "CalTPA_Commission Item 3D - June 2017" findings from the field test are presented in "CalTPA_Commission Item 2C - August 2018." The August 2018 Commission Item discusses how the CalTPA is grounded in the Universal Design for Learning theory. Because 2018–19 is the first operational year of the revised CalTPA, it is likely too soon to expect the model sponsor to have conducted extensive studies at this point. The model sponsor could use findings from the present comparability study to support intended use interpretations.
7.7	JS 2.13 The standard error of measurement, both overall and conditional (if reported), is provided in units of each reported score.	NA	Some small and/or specialized assessments cannot be expected to provide data that typically come from larger, more traditional assessment programs. Related to this point, with respect to reliability the Joint Standards state, "… there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment (p. 41)." The CalTPA model.

(continued)

Table 4.15. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
7.7	(continued)		sponsor does not provide conditional standard errors of measurement; however, see the rationales for Joint Standards 3.8, 4.18, 4.20, and 6.9 for a discussion of how CalTPA addresses scorer training, calibration, and monitoring of scoring accuracy. Moreover, CalTPA double scores all Cycle 1 and Cycle 2 portfolios for which the total score falls below the passing standard. By double scoring all submissions that are at or around the 'cut score' and all submissions that have more than one rubric score of '1' (see CalTPA QMP), CalTPA has a built-in safety net for addressing the classification accuracy of pass/fail decisions
7.8	JS 3.8 [The model sponsor] collects and reports evidence of the validity of constructed response score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores.	5	CalTPA collects and analyzes scoring statistics by gender, race/ethnicity, and languages spoken (i.e., English only, English and one or more other languages, one or more languages other than English), and setting/urbanicity of teaching placement.
7.9	JS 4.22 [The model sponsor] specifies the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.	5	CalTPA is a criterion-referenced assessment. CalTPA convened twenty-one content experts in May 2019 to recommend the passing standard based on discussion of necessary and acceptable levels of proficiency on the part of entry-level teachers. The "briefing book" method was used. The CalTPA Performance Assessment Overview (Version 02) document states that the CalTPA is one of multiple measures to inform candidate preparedness. It goes on to state that the CalTPA is intended to provide both a formal assessment of candidate ability and a framework of performance-based guidance to inform candidate preparation and continued professional growth. Furthermore, it states that feedback provided at the completion of each cycle is intended to facilitate preparation for the subsequent assessment cycle and that data is shared with institutions to assist them in making program improvements and to guide induction programs as they work with new teachers to individualize learning plans. The CalTPA is intended to provide authentic evidence of teaching ability and student learning experienced during clinical practice. The Candidate Score Report states that the Results Report "is for your records only" and that, "This assessment was not designed to compare your performance to that of other candidates. Your score is used to compare your performance to the performance level set by.

(continued)

Table 4.15. (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
7.9	(continued)		the Commission on Teacher Credentialing." Score reports include a section called "Understanding Your CalTPA Assessment Results Report," which is included in this report in Appendix 4.R. Interpretation of scores is provided in the subsections: Rubric Performance Summary and Cycle Performance Summary
7.10	JS 5.0 [The model sponsor] documents evidence of fairness, reliability, and validity of test scores for their proposed use.	4	During the 2018–19 school year, CalTPA representatives calculated demographic and total score descriptive performance statistics (number, percent, mean, standard deviation, and median, minimum, maximum) by gender, ethnicity, language (e.g., English Only or multilingual), and setting (i.e., urbanicity) subgroups. See also rationales for 4.9, 4.10, 4.13, 4.15, 4.16, 4.20 to 4.24. Additional evidence of fairness, reliability, and validity of test scores for their proposed use is not available at this early stage (i.e., first operational year).
7.11	JS 5.0 Test scores are derived in a way that supports the interpretations of test scores for the proposed uses of tests.	5	CalTPA tasks and rubrics are organized around the Plan, Teach and Assess, Reflect, Apply teaching cycle. Each rubric specifies the relevant TPEs being assessed, and the two tasks (Cycle 1 and Cycle 2), together, provide the candidate an opportunity to demonstrate mastery of the TPEs. Scores are provided on a rubric-by-rubric and total score basis.
7.12	JS 6.10 Reports and feedback are designed to support valid interpretations and use, and minimize potential negative consequences.	4	CalTPA score reports provide scores for each rubric and overall score for each Cycle (see Appendix 4.R). CalTPA score reports include an overall Pass/Fail status on the report. Pass/Fail decisions are made based on the overall score, although candidates fail if they get more than one score of 1 on the rubrics. The total score draws from a larger set of performances to create a more reliable measure on which to base pass/fail decisions. To help support valid interpretations and to minimize potential negative consequences, CalTPA should consider including guidance in the score reports that rubric level scores are provided for formative purposes and that the CalTPA total score should be used in conjunction with other measures of performance to determine a candidate's preparedness for beginning teaching.

Note. NA = Not applicable; CR = Cannot rate at this time.

Table 4.16. Claim 8 Ratings on the Assessment Design and Joint Standards for CalTPA

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
8.1	ADS 1(a) Collectively, the tasks and rubrics in the assessment address key aspects of the six major domains of the TPEs.	5	The TPE elements (i.e., key aspects) associated with each rubric are mapped to each rubric for Cycle 1 and Cycle 2 in the candidate performance assessment guides.
8.2	ADS 2(a) In relation to the key aspects of the major domains of the TPEs, the pedagogical assessment tasks, rubrics, and the associated directions to candidates are designed to yield enough valid evidence for an overall judgment of each candidate's pedagogical qualifications for a Preliminary Teaching Credential as one part of the requirements for the credential.	5	CalTPA includes complex performance tasks that require candidates to perform tasks and activities aligned with the elements (key aspects) of the six TPE domains. Multiple robust rubrics, which are mapped to key aspects of the TPE domains, are utilized to evaluate the submissions; candidates are required to provide multiple pieces of evidence for each rubric. Performance Assessment Guides inform candidates of the TPE elements (i.e., key aspects) measured by each CalTPA rubric.
8.3	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	See also rationale for 7.4 above in Table 4.15. The candidate score reports under "Rubric Performance Summary" states that "this information may help you identify your relative strengths and areas for improvement." To inform program quality and effectiveness, CalTPA provides programs with access to ResultsAnalyzer®. Analysis of candidate performance across candidate- and aggregate- level data with state averages are provided to programs through a propriety program, ResultsAnalyzer®. The revision of CalTPA was spurred in part because the model's new analytic rubrics provide detailed feedback to help programs determine how to better align their instruction and curriculum to the TPEs.
8.4	JS 6.10 [Score report] interpretations describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.	4	CalTPA score reports provide in simple language what the test covers (see Appendix 4.R). There is a statement in the score report that, "This information may help you identify your relative strengths and areas for improvement." However, there is no statement that performance feedback at the end of one cycle is intended to facilitate preparation for the subsequent cycle. Furthermore, there is no statement that CalTPA is one of multiple measures to inform candidate preparedness, nor that data is shared with institutions to assist them in making program improvements and to guide induction programs as they work with new teachers to individualize learning plans. These intended uses are covered in the CalTPA Performance Assessment Guide Overview (Version 02), but not in score reports.

(continued)

Table 4.16 (Continued)

#	Assessment Design/Joint Standard	CalTPA Rating	Rationale for CalTPA Rating
8.5	JS 6.10 Score precision [is] depicted by error bands or likely score ranges, showing the standard error of measurement.	NA	Some small and/or specialized assessments cannot be expected to provide data that typically come from larger, more traditional assessment programs. Related to this point, with respect to reliability the Joint Standards state, "… there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment (p. 41)." The CalTPA model sponsor does not provide conditional standard errors of measurement; however, see the rationales for Joint Standards 3.8, 4.18, 4.20, and 6.9 for a discussion of how CalTPA addresses scorer training, calibration, and monitoring of scoring accuracy. Moreover, CalTPA double scores all Cycle 1 and Cycle 2 portfolios for which the total score falls below the passing standard. By double scoring all submissions that are at or around the 'cut score' and all submissions that have more than one rubric score of '1' (see CalTPA QMP), CalTPA has a built-in safety net for addressing the classification accuracy of pass/fail decisions.
8.6	JS 6.10 The interpretive materials prepared by the [model sponsor] address common misuses or misinterpretations.	5	The Candidate Score Report states that the Results Report "is for your records only" and that, "This assessment was not designed to compare your performance to that of other candidates. Your score is used to compare your performance to the performance level set by the Commission on Teacher Credentialing." It may be helpful to provide guidance about if and how scores should be listed on teaching resumes or used elsewhere outside of the degree program by teacher candidates who pass.

Comparison of the Strength of Evidence for the Standards across TPA Models

In Tables 4.17 – 4.20, we provide a summary of the ratings on each ADS/*Joint Standard* evaluative statement for all three assessment models by each claim (without the rationale for each rating).

Across all ADS and *Joint Standards* that are applicable to Claim 3 ("*The **scoring rubrics** for each TPA model are sufficiently clear and detailed to ensure that trained raters can accurately and consistently score candidate submissions."*), the average ratings for Claim 3 are 4.14, 4.57, and 4.86, respectively, for FAST, edTPA, and CalTPA. This indicates that, overall, the evidence demonstrates adherence to all or most aspects of the ADS and *Joint Standards*.

Table 4.17. Comparison of Ratings on Claim 3 Assessment Design/Joint Standard Evaluative Statements across TPA Models

#	Standards	FAST Rating	edTPA Rating	CalTPA Rating
3.1	ADS 1(a) The assessment [includes] multi-level scoring rubrics that are clearly related to the TPEs that the task measures.	5	4	5
3.2	ADS 1(c) The model sponsor defines scoring rubrics so candidates for credentials can earn acceptable scores on the Teaching Performance Assessment with the use of different content-specific pedagogical practices that support implementation of the TK-12 content standards and curriculum frameworks.	5	4	5
3.3	ADS 1(h) The model sponsor develops scoring rubrics that focus primarily on teaching performance.	5	5	5
3.4	ADS 1(h) The model sponsor develops scoring rubrics 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	5	5
3.5	JS 4.18 Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.	3	5	5
3.6	JS 4.18 [Scoring criteria are] presented by the [model sponsor] with sufficient detail and clarity to maximize the accuracy of scoring.	3	5	5
3.7	JS 4.18 [The model sponsor provides] multiple examples of responses at each score level for use in training scorers and monitoring scoring consistency. [These] are typically added to scoring specifications during item development and tryouts.	3	4	4
	Average	4.14	4.57	4.86

As presented in Table 4.18, the average ratings for the Standards relevant to Claim 4 ("*For each TPA model, there is a comparable, comprehensive process to select, establish calibration, and train the assessors who score candidate submissions."*) were 3.70, 4.96, and 5.00, respectively, for FAST, edTPA, and CalTPA. This indicates that, overall, the evidence for FAST demonstrates adherence to some (but not all) key aspects of the ADS and *Joint Standards* and that all or nearly all aspects of the ADS and *Joint Standards* were addressed by the evidence for edTPA and CalTPA.

Table 4.18. Comparison of Ratings on Claim 4 Assessment Design/Joint Standard Evaluative Statements across TPA Models

#	Standards	FAST Rating	edTPA Rating	CalTPA Rating
4.1	ADS 1(g) The TPA model sponsor [provides] materials appropriate for use by [assessors] to become familiar with the design of the TPA model, the candidate tasks, the scoring rubrics, [and scoring processes].	5	5	5
4.2	ADS 1(h) The model sponsor develops 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	5	5	5
4.3	ADS 2(c) The assessor training program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and multi-level scoring rubrics.	5	4	5
4.4	ADS 2(c) The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy and calibration in relation to the scoring rubrics associated with the task.	4	5	5
4.5	ADS 2(c) The model sponsor uses only assessors who successfully calibrate during the required TPA model assessor training sequence.	3	5	5
4.6	ADS 2(c) When new pedagogical tasks and scoring rubrics are incorporated into the assessment, the model sponsor provides additional training to the assessors, as needed.	5	NA	5
4.7	ADS 2(e) All approved models must include a local scoring option in which the assessors of candidate responses are program faculty and/or other individuals identified by the program who meet the model sponsor's assessor selection criteria. These local assessors are trained and calibrated by the model sponsor, and whose scoring work is facilitated, and their scoring results are facilitated and reviewed by the model sponsor.	5	5	CR
4.8	ADS 2(e) The model sponsor must provide an annual audit process that documents that local scoring outcomes are consistent and reliable within the model for candidates across the range of programs using local scoring and informs the Commission where inconsistencies in local scoring outcomes are identified. If inconsistencies are identified, the sponsor must provide a plan to the CTC for how it will address and resolve the scoring inconsistencies both for the current scoring results and for future scoring of the TPA.	5	5	CR
4.9	ADS 2(e) The model sponsor provides a detailed plan for establishing and maintaining scorer accuracy and inter-rater reliability during field testing and operational administration of the assessment.	3	5	5
4.10	ADS 2(e) Scoring process conducted by the model sponsor to assure the reliability and validity of candidate outcomes on the assessment may include, for example, regular auditing, selective back reading, and double scoring of candidate responses near the cut score by the qualified, calibrated scorers trained by the model sponsor.	4	5	5

(continued)

Table 4.18 (Continued)

#	Standards	FAST Rating	edTPA Rating	CalTPA Rating
4.11	JS 3.0 All [scoring procedures steps] are designed in such a manner as to minimize construct-irrelevant variance.	3	5	5
4.12	JS 3.4 Test takers receive comparable treatment during the [scoring process]. Those responsible for testing adhere to standardized scoring protocols so that test scores will reflect the construct(s) being assessed and will not be unduly influenced by idiosyncrasies in the testing process.	3	5	5
4.13	JS 3.8 Adequate training and calibration of scorers is carried out and monitored throughout the scoring process to support the consistency of scorers' ratings for individuals from relevant subgroups. Where sample sizes permit, the precision and accuracy of scores for relevant subgroups also is calculated.	3	5	5
4.14	JS 3.8 For human scoring, scoring procedures [are] designed with the intent that the scores reflect the examinee's standing relative to the tested construct(s) and are not influenced by the perceptions and personal predispositions of the scorers.	4	5	5
4.15	JS 4.20 Specifications should describe processes for assessing scorer consistency and potential drift over time in raters' scoring.	3	5	5
4.16	JS 4.20 The basis for determining scoring consistency (e.g., percentage of exact agreement, percentage within one score point, or some other index of agreement) are indicated.	5	5	5
4.17	JS 4.20 The process for selecting, training, qualifying, and monitoring scorers is specified by the [model sponsor].	3	5	5
4.18	JS 4.20 To the extent possible, scoring processes and materials anticipate issues that may arise during scoring.	3	5	5
4.19	JS 6.9 [The model sponsor has] procedures in place to monitor consistency of scoring across administrations (e.g., year-to-year comparability).	NA	5	NA
4.20	JS 6.9 [The model sponsor] appropriately retrains, rescores, and dismisses some scorers, and/or reexamines the scoring rubrics or programs based on inaccurate or inconsistent scoring	3	5	5
4.21	JS 6.9 Analyses monitor possible effects on scoring accuracy of variables such as scorer, task, time or day of scoring, scoring trainer, scorer pairing, and so on, to inform appropriate corrective or preventative actions.	2	5	5
4.22	JS 6.9 Consistency in applying scoring criteria is checked by independently rescoring randomly selected test responses.	4	5	5
4.23	JS 6.9 Periodic checks of the statistical properties (e.g., means, standard deviations, percentage of agreement with scores previously determined to be accurate) of scores assigned by individual scorers during a scoring session are used to provide feedback for the scorers, helping them to maintain scoring standards.	1	5	5
4.24	JS 6.9 Those responsible for scoring document the procedures followed for scoring, procedures followed for quality assurance of that scoring, the results of the quality assurance, and any unusual circumstances.	4	5	5
	Average	3.70	4.96	5.00

Note. CR= Cannot rate at this time.; NA= Not applicable.

As presented in Table 4.19, the average ratings for the Standards relevant to Claim 7 ("*For each TPA model, the score reports, candidate-level and program-level, provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used."*) were 4.36, 4.64, and 4.70, respectively, for FAST, edTPA, and CalTPA. This indicates that, overall, the evidence demonstrates adherence to most or all aspects of the ADS and *Joint Standards* for all three models.

Table 4.19. Comparison of Ratings on Claim 7 Assessment Design/Joint Standard Evaluative Statements across TPA Models

#	Standards	FAST Rating	edTPA Rating	CalTPA Rating
7.1	ADS 2(g) The model sponsor conducting scoring for the program provides results on the TPA to the individual candidate based on performance relative to TPE domains and/or to the specific scoring rubrics within a maximum of three weeks following candidate submission of completed TPA responses.	5	4	5
7.2	ADS 2(g) The model sponsor follows the timelines established with programs using a local scoring option for providing scoring results.	5	5	CR
7.3	ADS 2(g) The model sponsor provides results to programs based on both individual and aggregated data relating to candidate performance relative to the rubrics and/or domains of the TPEs.	5	4	5
7.4	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	4	4
7.5	JS 1.1 The [model sponsor] specifies in clear language the contexts in which test scores are to be employed.	4	5	5
7.6	JS 1.2 A summary of the evidence and theory bearing on the intended interpretation is presented for each intended interpretation of test scores for a given use. Evidence may come from studies conducted locally, in the setting where the test is to be used; from specific prior studies; or from comprehensive statistical syntheses of available studies meeting clearly specified study quality criteria. No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test score interpretation for a given use determine the value of a particular kind of evidence.	4	5	5
7.7	JS 2.13 The standard error of measurement, both overall and conditional (if reported), is provided in units of each reported score.	NA	NA	NA
7.8	JS 3.8 [The model sponsor] collects and reports evidence of the validity of constructed response score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores.	5	5	5
7.9	JS 4.22 [The model sponsor] specifies the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.	5	5	5
7.10	JS 5.0 [The model sponsor] documents evidence of fairness, reliability, and validity of test scores for their proposed use.	4	5	4
7.11	JS 5.0 Test scores are derived in a way that supports the interpretations of test scores for the proposed uses of tests.	4	5	5
7.12	JS 6.10 Reports and feedback are designed to support valid interpretations and use, and minimize potential negative consequences.	3	4	4
	Average	4.36	4.64	4.70

Note. CR = Cannot rate at this time; NA = Not applicable.

As presented in Table 4.20, the average ratings for the Standards relevant to Claim 8 ("*The scoring rubrics and score reports provide diagnostic information on candidates and on programs such that the strengths and weaknesses of each can be identified.*") were 4.20, 4.20, and 4.60, respectively, for FAST, edTPA, and CalTPA. This indicates that, overall, the evidence demonstrates adherence to most or all aspects of the ADS and *Joint Standards* for all three models.

Table 4.20. Comparison of Ratings on Claim 8 Assessment Design/Joint Standard evaluative statements across TPA Models

#	Standards	FAST Rating	edTPA Rating	CalTPA Rating
8.1	ADS 1(a) Collectively, the tasks and rubrics in the assessment address key aspects of the six major domains of the TPEs.	5	4	5
8.2	ADS 2(a) In relation to the key aspects of the major domains of the TPEs, the pedagogical assessment tasks, rubrics, and the associated directions to candidates are designed to yield enough valid evidence for an overall judgment of each candidate's pedagogical qualifications for a Preliminary Teaching Credential as one part of the requirements for the credential.	5	4	5
8.3	JS 1.1 The [model sponsor] sets forth clearly how test scores are intended to be interpreted and consequently used.	4	4	4
8.4	JS 6.10 [Score report] interpretations describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.	3	4	4
8.5	JS 6.10 Score precision [is] depicted by error bands or likely score ranges, showing the standard error of measurement.	NA	NA	NA
8.6	JS 6.10 The interpretive materials prepared by the [model sponsor] address common misuses or misinterpretations.	4	5	5
	Average	4.20	4.20	4.60

Discussion

In Activity 4, we investigated the extent to which the following claims are supported by the available documentation and evidence pertaining to scoring rubrics, scorer training, and score reports:

- **Claim 3:** The scoring rubrics for each TPA model are sufficiently clear and detailed to ensure that trained raters can accurately and consistently score candidate submissions.
- **Claim 4:** For each TPA model, there is a comparable, comprehensive process to select, train, and establish calibration of the assessors who score candidate submissions.
- **Claim 7:** For each TPA model, the score reports (candidate-level and program-level) provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used.
- **Claim 8:** The rubrics and score reports provide diagnostic information on candidates and on programs such that the strengths and weaknesses of each can be identified.

To conduct this study, we (a) reviewed TPA model documentation, (b) observed scoring-related trainings and/or accessed and interfaced with applicable TPA model scorer training platforms, and (c) interviewed model representatives involved with scoring. We then rated each TPA model on a set of evaluative criteria developed by first locating applicable ADS and *Joint Standards* that aligned to each claim being investigated and then parsing the Standards into evaluative statements so that a single issue was defined in each evaluative statement (the evaluative statements are what is presented in Tables 4.5 – 4.16). In this section, we discuss the findings by claim.

Claim 3: The scoring rubrics for each TPA model are sufficiently clear and detailed to ensure that trained raters can accurately and consistently score candidate submissions.

The evaluative statements (i.e., Standards parsed into a single issue) were considered in relation to Claim 3. Designing scoring rubrics and criteria for complex performances like those required by beginning teachers requires considerable coordination across many stakeholders to ensure a cohesive approach to candidate assessment. It typically also requires several iterations to ensure a linkage among the content domain, performance tasks, and rubrics. Candidate responses that cover a wide range of competency should be evaluated to determine the extent to which scoring rubric criteria reflect the components displayed in candidate work. Scoring criteria may then be modified and/or the task may be redesigned to ensure it assesses the intended knowledge, skills and abilities as described in the TPEs.

After review of the available information for FAST's, edTPA's, and CalTPA's rubrics, we found that, overall, they are sufficiently clear and detailed to ensure that trained raters can accurately and consistently score candidate submissions. To be specific, we found the rubrics of each of the three TPAs:

- Focus on teaching performance,
- Link to the TPEs (although linkage to TPEs is clearer/more apparent for FAST and CalTPA than edTPA because FAST and CalTPA include those linkages in the candidate manual/guides),
- Allow candidate responses to cover a wide range of competency while still matching scoring criteria,
- Contain score level descriptors that clearly delineate what a candidate must know or do to earn each score level, and
- Minimize candidate factors, qualities, or characteristics not likely to affect job effectiveness and/or student learning.

The most notable differences in the rubrics across the models include (a) the number of rubrics, (b) the number of score levels, (c) subject specificity of rubrics, (d) format of rubrics, and (e) the clarity/transparency of the linkage of rubrics to TPEs. First, the CalTPA and edTPA models have more rubrics than FAST (17 for CalTPA and 15 or 18 for edTPA,[29] whereas FAST has 10). Second, FAST uses four score levels with labels on each level ("Does Not Meet Expectations" through "Exceeds Expectations"), whereas edTPA and CalTPA use five levels that are simply labeled: Level 1, 2, 3, 4, 5. Third, edTPA and CalTPA have more rubric language tailored to

[29] There are 15 rubrics for single subject credential areas and 18 rubrics for the multiple subject credential.

specific subject areas, whereas FAST has just one rubric (and one candidate manual) that is applied to all credential areas.

In addition, the format of the FAST rubrics is distinct from the edTPA and CalTPA rubrics in that the FAST rubrics contain aspects of analytic and holistic scoring, whereas edTPA and CalTPA rubrics are analytic. For FAST, each of the 10 rubrics contains 2–3 indicators. Each indicator contains multiple behavioral descriptors for each score level. For example, for the Reflection rubric there are three indicators: (a) subject specific pedagogy, (b) applying knowledge of students, and (c) student engagement. Each of those indicators contain multiple behavioral descriptors for each score level. Scorers are instructed to determine a level rating for each indicator, and then to use their judgment to come up with an overall rating for the rubric. Currently, there is no clear guidance on how FAST scorers should collapse over the indicators to derive an overall rubric rating. One advantage of the FAST rubric is that it provides candidates with a rich set of information to guide their portfolio submissions. A disadvantage, however, is that FAST rubrics likely require more effort per rubric by scorers to determine where a candidate submission falls as compared to edTPA and CalTPA, which simply require scorers to identify one score level on each rubric. In addition, given that FAST rubric scores are derived by collapsing over indicator ratings, it may be less apparent to FAST candidates why they received a particular score on a rubric. However, this may be offset to some extent by the fact that FAST is a small, local program and candidates likely have more opportunity to obtain detailed feedback directly from their coaches.

Finally, the edTPA rubrics differ from the FAST and CalTPA rubrics in that the linkage between rubrics and TPEs are readily available to FAST and CalTPA candidates and programs through the candidate manual (FAST)/Performance Assessment Guides (CalTPA). Moreover, rubric level scores on score reports are also linked to TPE elements for one of the FAST components (TSP). The edTPA model sponsors have provided documentation to the Commission in its 556-page Transition Plan that maps edTPA rubrics to TPEs. However, the linkage between edTPA rubrics (and tasks) and TPEs is not readily available to candidates or programs.

In conclusion, the results of our investigation show that all three TPA models are reasonably comparable in meeting the ADS and industry standards (*Joint Standards)* related to scoring rubrics. Related to the current TPA model implementation, we recommend that FAST develop written guidance for assessor rubric use, especially how to weight indicators within each rubric. A short guideline for determining how to weigh the importance of individual indicators for rubric ratings that is provided at future scorer training sessions could improve the reliability of FAST scores. We also recommend that edTPA consider making the linkage between TPE elements and edTPA rubrics and tasks readily available to candidates and programs, perhaps via a supplemental linkage document that can be made available to candidates and programs. Finally, we recommend that all TPA models ensure all levels of their rating scales are presented to scorers at training. Providing more examples of all rubric rating levels will improve assessor clarity at scorer training sessions and, if posted publicly, for candidates, program staff, and other stakeholders. Exemplars at the extremes of the scales (e.g., Level 1 and Levels 4 and 5) were noticeably underrepresented at all observed TPA training sessions and in all TPA training materials.

Claim 4: For each TPA model, there is a comparable, comprehensive process to select, establish calibration, and train the assessors who score candidate submissions

The results indicate that the scoring processes for all three models address key aspects of the ADS and *Joint Standards* relating to scorer training. However, edTPA and CalTPA currently have stronger processes and better documentation in place to ensure that scorers maintain the

calibration attained during training. First, related to scorer selection, edTPA and CalTPA recruit scorers with subject knowledge and a recent background in teaching or teacher education. The main difference between edTPA and CalTPA scorer selection is that CalTPA has the additional criterion that their assessors must have experience or a credential from the state of California. FAST, unlike edTPA and CalTPA, does not invite TK-12 teachers or anyone outside of the program to score. As a local, self-contained assessment program, it exclusively uses its own, Fresno State teacher education faculty and staff members. Furthermore, for FAST, only scorers who have taught in that subject at the university or have a credential to teach it in a K-12 school may score. A potential limitation of FAST is that Single Subject teacher supervisors score the candidates whom they supervise because there aren't enough scorers with subject expertise to score teacher candidates who aren't on their roster. While we don't have any direct recommendation related to this, we note that it may be difficult for a supervisor to provide an objective evaluation of a teacher candidate because of their previous knowledge and observation of the teacher candidate. Further, the supervisor might be reluctant to score strictly because the candidate will know who provided the low or non-passing score. On the other hand, one might argue that having more knowledge of the candidate may allow the scorer to provide more useful and actionable feedback for candidates and result in better and even more accurate scores.

Assessment Design Standard 2(c) requires that, "*The assessor training program demonstrate convincingly that prospective and continuing assessors gain a deep understanding of the TPEs . . .*" Because FAST and CalTPA use California educators as scorers these models have an inherent advantage over edTPA, which uses a national sample of scorers, for ensuring this requirement is met. Moreover, both FAST's and CalTPA's scoring rubrics are mapped to the TPE elements and this mapping is readily available to all stakeholders through the Candidate Manual (FAST)/Performance Assessment Guides (CalTPA). Thus, as FAST and CalTPA assessors are deepening their understanding of the scoring rubrics during training, they are simultaneously deepening their understanding of the TPE elements which are mapped to those rubrics. In contrast, edTPA uses a national sample of scorers and the edTPA scoring rubrics as they appear in the edTPA Handbooks and score reports are not mapped to the TPE elements. The edTPA model sponsor does provide a supplemental handout to edTPA scorers scoring California submissions called, "Deep understanding of the TPEs." Although the extent to which the provision of this supplemental handout contributes to edTPA scorers having a "deep understanding of the TPEs" is unknown.

Another notable difference among the models is that edTPA scorer training and calibration is conducted online, whereas FAST and CalTPA conduct in-person scorer training, although CalTPA scorers access practice and qualifying portfolios using the online ePEN platform.

There are commonalities in calibration procedures across all three models, but also some notable differences. All models require "practice" scoring on exemplar portfolios followed by feedback and discussion. Then, scorers score a qualifying portfolio. The edTPA requires scorers to calibrate on two qualifying portfolios, whereas FAST and CalTPA only require calibration on one.

All models require scorers to meet a minimum performance threshold prior to "qualifying" to score. FAST's minimum performance threshold is more stringent than edTPA's and CalTPA's performance threshold in that FAST does not allow for any non-adjacent scores, whereas both edTPA and CalTPA allow for one non-adjacent score. However, this more stringent criterion for FAST is offset by the fact that FAST uses a 4-point scoring rubric, and edTPA and CalTPA use a 5-point scoring rubric.

A notable difference between calibration procedures for FAST and those for edTPA and CalTPA is that in 2018–19, returning scorers for FAST were not required to re-calibrate, although they did attend training sessions to discuss the revisions that were made to the rubrics following field test. We recommend FAST require all its scorers, including returning scorers, to re-establish calibration on a qualifying portfolio, especially when revisions, even minor ones, are made to rubrics and/or tasks.

Another notable difference between FAST and the other two models are FAST's procedures for monitoring scorer consistency. FAST's small size limits its staff's ability to *monitor* calibration, inter-rater reliability, and retrain assessors *during* the scoring window compared to edTPA and CalTPA, which have many more scorers and use electronic scoring software. edTPA and CalTPA both have scoring systems with built-in features that monitor scorer performance automatically and in real time. FAST, which uses paper scoring, cannot. FAST examines inter-rater reliability *after* scoring takes place and all scores have been reported to teacher candidates; this does not allow opportunity to make corrective changes or provide real-time feedback to scorers. We recommend that FAST incorporate a calibration exercise near the middle of each scoring window to ensure all scorers are still scoring consistently/are calibrated. This exercise, even if brief, would help ensure that scorers that have drifted can recalibrate to the model's rating levels and expectations.

In conclusion, training and calibration are thorough for all edTPA and CalTPA scorers and FAST scorers who are new. The edTPA and CalTPA trainings require parts of several days. FAST requires less time; however, the scorers are already familiar with the tasks required for the assessment. CalTPA and FAST both conduct their training in person. The edTPA training sessions are conducted remotely via online tutorials. CalTPA and edTPA both require re-calibration of all returning scorers, whereas FAST did not in 2018–19. Across the TPA models, we found the training of each:

- provides appropriate materials (e.g., example submissions) to scorers to help them develop knowledge of the candidate tasks and rubrics,
- limits bias of scorers via training discussion and/or verbal reminders, and
- provides comparable treatment to each candidate.

Again, we recommend FAST require returning scorers to recalibrate. Lack of recalibration in 2018–19 was concerning because of the changes made to descriptions of the indicator rating levels in the rubrics following field test administration in 2017-2018.

Claim 7: For each TPA model, the score reports (candidate-level and program-level) provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used.

The TPA models should provide teacher candidates with accurate and useful score reports that inform candidates of whether they met a requirement for preparedness for beginning teaching. Aggregated, the data from these assessments should also inform institutions and their stakeholders of program quality and effectiveness. Ideally, stakeholders, especially teacher candidates, should find the assessment results easy to access, easy to interpret, and able to support uses aligned to the purposes of the assessment.

While the TPA models generally compared very well and demonstrated many best practices related to score reporting, there are some notable differences and some areas noted for improvement. These topics are discussed below.

First, CalTPA's score reporting to candidates and programs appears similar to edTPA's. Both provide similar rubric level scores to teacher candidates along with total scores, although they differ in that CalTPA score reports directly state whether the candidate passed the assessment, whereas edTPA score reports direct candidates to a website to look up the passing requirement for their state. At the program level, both edTPA and CalTPA provide files with teacher candidate scores and access to the same online platform, ResultsAnalyzer®, to conduct data exploration. FAST score reports differ from edTPA and CalTPA in that the FAST score reports only include rubric level scores; there is no total score, nor indication of pass/fail status included on the FAST score reports. Being a local program, FAST has score results once assessors finish their scoring assignments. Raw data is provided to program supervisors as it becomes available.

Another difference among the models is the information included in score reports on intended use of scores. All models include documentation in their supporting materials that states how scores should be interpreted and used; however, they differ regarding inclusion of that information on score reports. On one end, the FAST model does not include in its score reports information on how scores should be interpreted and used. In contrast, both edTPA and CalTPA include guidance in their score reports that scores are used to compare candidates' knowledge and skills/performance to the requirements set by their state/Commission. In addition, both edTPA and CalTPA include guidance in score reports that scores are not to be used to compare one's performance to that of other candidates. The CalTPA score report also includes guidance that the rubric level scores "may help you identify your relative strengths and areas for improvement." Because total scores are based on a larger sample of performance than individual rubric scores, and thus are more reliable than rubric level scores, it may be helpful to include guidance in score reports that rubric level scores should be used formatively for identifying strengths and weaknesses and that total scores should be used for making pass/fail decisions. It's worth noting that FAST does not report total scores and that if candidates receive a '1' (Does Not Meet Expectations) on any rubric, then they do not pass. These candidates are contacted via email and notified of the sections (i.e., rubrics) on which they received a non-passing score. They are informed that they have the opportunity to redo those section(s) and to contact the coordinator/supervisor to make an appointment to discuss what they will need to do to revise that section(s). During an interview with the FAST Coordinator, the model sponsor shared that they downplay the scores and focus a lot more on what candidates are learning by going through the process. Interestingly, 100% of FAST candidates passed the assessment in 2018–19 (including retakes). This reinforces the sentiment expressed by the FAST Coordinator—that FAST is functioning primarily as a formative tool.

Finally, it is worth noting that *Joint Standard* 2.13 states that "the standard error of measurement, both overall and conditional (if reported), [should be] provided in units of each reported score." After much consideration and deliberation, we ultimately decided to rate this standard as "not applicable" for inclusion in score reports. Related to this point, the *Joint Standards* state, "… *there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment* (p. 41)." It is important to note that while precision of measurement information is not included on the TPA score reports all models

have procedures in place to ensure the classification accuracy of pass/fail decisions (i.e., for FAST all non-passing rubric scores are double scored and for edTPA and CalTPA all submissions whose total scores are at or around the cut score are double scored). This added scrutiny of scores near the cut score serves the purpose of limiting classification inaccuracy.

In conclusion, across the three TPA models, we found that score reports:

- Provide teacher candidates with scores at the rubric level, and, for edTPA and CalTPA, at the total score level,
- Release results on a timeline that adheres to the ADS,
- Provide programs with individual and aggregated data relating to candidate performance, and
- Specify the criterion (i.e., the rubric language) used to derive each score.

All three models may want to consider including additional guidance on their score reports regarding appropriate score use. Again, all three models include this information in their reference materials, but not directly on score reports. For example, none of the models include guidance on their score reports that scores should be used in conjunction with other measures to determine a candidate's preparedness for beginning teaching. Moreover, CalTPA is currently the only model that includes guidance in its score reports that the rubric level scores (as opposed to total scores) may help candidates identify their relative strengths and areas for improvement. Given that rubric scores are essentially subscores (i.e., based on a smaller sample of performance than total scores) and therefore are inherently less reliable than total scores, it is important that rubric level scores include this caution.

Claim 8: The rubrics and score reports provide diagnostic information on candidates and on programs such that the strengths and weaknesses of each can be identified.

To evaluate Claim 8, we extended the evaluation of score reports conducted for Claim 7 by specifically focusing on the diagnostic information on candidates and programs provided by each TPA model. As discussed under the discussion for Claim 7, all models report rubric level scores, although only CalTPA score reports include guidance that rubric level scores "may help you identify your relative strengths and areas of improvement." The FAST and edTPA models may want to consider including similar guidance in their score reports, although as noted above, we recommend that models convey that rubric scores and overall scores be used in conjunction with other information to make determinations about a candidate's readiness for beginning teaching.

Aside from score reports, the models include guidance in other resource materials on how candidates can use their scores diagnostically. The FAST Candidate Manual opens with a letter to candidates. In this letter, it states that, "A history of your scores will be available to you through Tk20 for sharing with your professional induction program supervisor as you see fit." This guidance suggests that candidates can use FAST scores to inform their continued professional development. The CalTPA Performance Assessment Overview (Version 02) document states that CalTPA is intended to provide both a formal assessment of candidate ability and a framework of performance-based guidance to inform candidate preparation and continued professional growth. Furthermore, it states that feedback provided at the completion of each cycle is intended to facilitate preparation for the subsequent assessment cycle. Finally, edTPA does include guidance in materials that serve as supplements to score reports, and

which discuss the learning associated with the outcomes of edTPA and how those formative experiences are guided by the structure and progression of the edTPA rubrics.[30] In terms of providing diagnostic information at the program-level, both edTPA and CalTPA provide programs access to a proprietary program called ResultsAnalyzer®, which can be used to generate custom views and reports on program performance. Per the documentation provided, ResultsAnalyzer® allows programs to generate customized test performance reports to inform state policy, research, and state accountability efforts. However, because we did not collect customized reports from preparation programs that use edTPA or CalTPA, the extent to which programs are using ResultsAnalyzer® for these purposes is unknown. The edTPA model sponsor also produces a biannual report that is available to programs that contains descriptive statistics about national, state, and program-specific populations, although, again, the extent to which programs are using these biannual reports to inform their own program's quality and effectiveness is unknown. The CalTPA Performance Assessment Overview (Version 02) document states that data is shared with institutions to assist them in making program improvements and to guide induction programs as they work with new teachers to individualize learning plans. For FAST, faculty receive score data for their credential area although it's not clear how this information is used. In 2018–19, the Commission did not require an annual report. This may be an impetus for the models, and FAST, in particular, to analyze program level results. In the meantime, we recommend that FAST compute basic descriptive analyses on credential areas to help inform program quality and effectiveness, which is a stated intended use of their score data.

Finally, the analytic nature of the rubrics themselves are useful for providing diagnostic information for candidates and programs, although as discussed above under Claim 3, the FAST rubric, with its multiple descriptors for each score level for each indicator (2–3 indicators per rubric), provides candidates with a rich set of information to guide their portfolio submissions, but there is less transparency in how scores were derived given that scorers are instructed to use their judgment in collapsing over indicator-level ratings to obtain an overall rubric-level score.

Conclusion

Overall, the findings from this study (i.e., Activity 4) indicate that there are more similarities than differences across models in topics related to scoring and score reporting, although the differences are notable. In summary, regarding Claim 3, all three models have clear and detailed scoring rubrics that help to ensure that trained scorers can accurately and consistently score candidate submissions, although clearer guidance to FAST scorers on how to weight indicator level ratings on each rubric may help to further strengthen scorer consistency. Furthermore, all three models use multi-level scoring rubrics that are clearly related to teaching performance expectations, although the linkage of scoring rubrics to TPEs are more readily available/transparent for FAST and CalTPA (i.e., via Candidate Manual/Guides) than for edTPA. Regarding Claim 4, all three models carefully select, train, and calibrate scorers, although in 2018–19 returning FAST scorers attended training, but were not required to re-calibrate. Also, edTPA and CalTPA have stronger procedures in place than FAST to monitor scorer consistency. Currently, interrater reliability analyses for FAST are only conducted after scoring is completed to demonstrate scoring reliability, not to identify and remediate scorers who may have drifted from the calibration standard. Regarding Claim 7, all models provide rubric level scores to candidates and programs in a timeframe that's consistent with ADS requirements. Unlike the edTPA and CalTPA models, FAST does not include the total score nor an indication of pass/fail status on the candidate score report (edTPA does not include the pass/fail status on the score report like CalTPA, but edTPA does include a weblink to where candidates can look

[30] For example, see https://www.edtpa.com/Content/Docs/GuidelinesForSupportingCandidates.pdf

up the passing standard for their state). Moreover, all models provide guidance on how candidate score information should be used, although the models differ with regard to the inclusion of that information on score reports. The FAST model does not include guidance on score reports about how scores should be used, but it does include that guidance in its Candidate Manual. Finally, regarding Claim 8, candidate score reports for all models are diagnostic in the sense that they report rubric level scores; however, only the CalTPA score report includes guidance that rubric level scores can be used to identify candidate strengths and weakness. Such guidance is not included on the edTPA and FAST score reports, although edTPA does include this guidance in materials that serve as supplements to the score reports, and the FAST Candidate Manual includes guidance that candidates can use their scores to inform continued professional growth. None of the models include guidance on their score reports that TPA scores should be used in conjunction with other measures for determining a candidate's readiness for beginning teaching, although all models do include such guidance in other supporting materials. In summary, FAST, edTPA, and CalTPA largely adhere to most ADS and *Joint Standards* related to scoring. On average, FAST was rated slightly lower than edTPA and CalTPA, but the lack of comparability may be balanced out, to some extent, by some of the unmeasured benefits its university supervisors achieve by being active in the credentialing process—something only a local program, like FAST, could achieve.

Chapter 5: Comparison of Standard Setting across TPA Models (Activity 5)

Wade Buckland & Andrea Sinclair

Introduction

The purpose of Activity 5 was to investigate Claim 5, which is:

> The standard-setting procedures used for each TPA model are sufficiently comparable and rigorous to ensure that the respective passing standards for each model accurately and consistently identify candidates possessing the requisite knowledge, skills, and abilities (KSAs) required to effectively teach the content area(s) authorized by the credential.

The determination of comparability across standard setting procedures is not straightforward when we consider the differences in the TPA models. From one perspective, if the TPA models generate similar proportions of passing candidates, we might assume that they are classifying candidates similarly. However, such a determination rests on many assumptions, such as the comparability of the knowledge, skills, and abilities (KSAs) of the candidate pools tested by each TPA and the comparability of emphasis across the *Teaching Performance Expectations* (TPEs) for the TPA models (which is the focus of Activity 2—see Chapter 2 in the Year 1 Report (Sinclair & Thacker, 2018).

Differences in standard-setting methods may influence whether the TPA models are comparably classifying candidates as "TPE-ready." For the passing standards to be comparable, they should reference similar definitions of minimally qualified candidates. They should have similar performance level descriptors and key differentiators for determining passing scores. The extent to which this is not the case represents a threat to the comparability of TPA models. The use of impact data can have a large effect on standard setters. Knowing the proportion of candidates likely to pass can cause panelists to shift their original ratings substantially. Often, panelists or facilitators will have a proportion passing, or an acceptable range for passing, in mind at the outset of standard-setting, which can also be a threat to the validity of the standard-setting.

To investigate Claim 5, HumRRO conducted observations of standard-setting procedures and reviews of standard-setting documentation. The comparison of the processes that the three models used to arrive at their present passing standards is not straightforward for two reasons. First, the standard-setting workshop for edTPA occurred July 1, 2014, which was prior to the start of this comparability investigation; thus, the HumRRO evaluation is unable to include an observation of the standard-setting workshop and instead relies exclusively on a documentation review. Second, the Passing Standard Workshop for FAST is a departure from common standard setting methods, such as Body of Work, Angoff, Briefing Book, etc. The standard setting method used by FAST most closely resembles the Dominant Profile Method in which panelists are asked to generate the policy (decision) rule that determines what scores across the tasks [in this case rubrics] are the minimum needed to "pass" (Hambleton, Jaeger, Plake, & Mills, 2000); however, in the case of the FAST standard setting, rather than asking panelists to first independently create their decision rule for passing, the panelists discussed the policy as a group, and the group discussion took the form of verifying through consensus discussion that the Level 2 descriptor ("Meets Expectations") for each rubric accurately describes a just sufficiently qualified candidate. Thus, the policy rule for FAST is that candidates must score a '2' on all rubrics for both components of FAST; there is no compensatory scoring and no impact data were considered, whereas such information was considered for the edTPA and CalTPA

standard settings. These differences across the models should be kept in mind when considering the findings presented in this chapter.

In the pages that follow, we describe the (a) the method we used to evaluate the standard-setting information, including a list of information sources available on standard-setting for each TPA model; (b) the findings for each model; and (c) comparisons among the three models.

Method

We used direct observation and documentation provided by the model sponsors to evaluate standard-setting procedures. Various sources of evidence and background information were available from the three models, including Transition Plans provided by the model sponsors, weblinks to source documents, manuals/handbooks/guides for the TPAs, PowerPoint slides presented at a TAC meeting, etc. (For a complete list of the available documentation see Appendix 1.A). In addition, email communications obtained from the model sponsors provided more recent information on standard-setting updates. We systematically reviewed all available relevant information and evaluated the models' standard setting processes based on industry standards. Our evaluation process is described in more detail below, following the description of evidence sources.

FAST Evidence Sources

The FAST model is comprised of two tasks, which the model sponsor refers to as the Site Visitation Project (SVP) and the Teaching Sample Project (TSP).

Observation. HumRRO conducted a site visit at Fresno State University School of Education on May 14, 2018 to observe the "Passing Standard Workshop" for the SVP task. The purpose of the workshop was to determine whether the existing rubric for a Level 2 ("Meets Expectations") reflects reasonable expectations for beginning teachers. HumRRO observers did not attend the Passing Standard Workshop for the TSP task, but communication from the model sponsor indicated that the TSP Passing Standard Workshop followed the same process.

The SVP Passing Standard Workshop began with a brainstorming activity during which the panelists, teacher preparation educators at Fresno State, listed the planning and teaching KSAs that they think a beginning teacher should have. The panelists stressed the importance of setting high expectations for the candidates and ensuring that they are aware of what it takes to become a highly qualified teacher. The workshop facilitator reminded the panelists that they should be thinking about a beginning teacher who may not have the characteristics of an experienced and successful teacher. To facilitate the differentiation between candidates who are not meeting the passing standard (i.e., a Level 1 on all rubrics) and those who are meeting the standard (i.e., a Level 2 or above), the facilitator suggested that panelists (a) provide examples of teacher performance that they themselves observed that were above or below the passing standard and (b) recall what specific practices were challenging for them as beginning teachers and why. The panelists were able to come up with behaviors characteristic of unsuccessful teacher candidates who they observed during site visits. The facilitator prompted the panelists to think about whether those actions by the teacher candidates would mean that they would fail the SVP task. Then, the facilitator directed their attention to how the skills of a passing teacher candidate differ from those of a non-passing teacher candidate.

As a final workshop activity, the panelists reviewed the three rubrics corresponding to the three parts of the SVP task (i.e., planning, implementation, reflection) for Level 1 (Does Not Meet

Expectations) and Level 2 (Meets Expectations). The purpose of the review was to evaluate whether the rubrics reflect reasonable expectations of what teacher candidates should know and be able to do, and whether the rubrics are worded clearly. They discussed some wording that appeared vague (e.g., "little or no understanding" and "typical student") and possible changes to the wording. Minor changes were made to the wording in the rubrics to improve clarity; the consensus among the panelists was that the rubrics for Level 2 reflect reasonable expectations of the KSAs required for beginning teachers. (The full SVP site visit report can be found in Appendix 5.A).[31]

Documentation. To supplement our observation of the SVP Passing Standard Workshop, we reviewed the documentation provided by the FAST model sponsor, including the SVP Brainstorm Table and the SVP rubric passing standard review document used in the SVP passing standard workshop (see Appendix 5.A for additional detail on the documents used during the passing standard workshop). In addition to the May 14, 2018 SVP standard setting activity, FAST held a Preliminary Passing Standard Workshop on August 4, 2017 and a TSP Passing Standard Workshop on May 29, 2018. The Preliminary Passing Standard Workshop included a review of the SVP and TPEs and a discussion of the rubrics by 12 participants (5 master teachers, 6 university coaches, and a Special Education faculty member; the group had a median of 1.5 years' experience with a variety of credential areas; five were men and seven were women; and seven representatives were white, two were Asian, and three were Latino). The TSP Passing Standard Workshop included six participants (all were university coaches, half represented a Single Subject credential area and half represented the Multiple Subject credential area, all participants were white, and one was male).

edTPA Evidence Sources

Observation. Because the standard-setting was conducted July 1, 2014, which was prior to the start of this comparability investigation, we were not able to conduct an observation.

Documentation. The standard-setting workshop and procedures for the edTPA are described in the edTPA Transition Plan and an agenda provided for an August 2014 Commission meeting. Per these documents, the standard-setting for edTPA was conducted using the Briefing Book Method (Haertel, 2005, 2008). Using this method, edTPA's standard-setting process was informed by a "briefing book," in which a compendium of relevant information to inform a standard-setting was compiled and made available to the participants in the standard-setting process. The briefing book described the design of edTPA and the goal of the standard-setting process. In addition, the briefing book contained evidence to (a) characterize the level of performance at different potential cut scores and (b) provide contextual information about the likely impact and appropriateness of different potential cut scores (e.g., passing rates).

The characterizations of performance at different potential edTPA cut scores served as performance standards corresponding to each cut score. Like other assessments using the briefing book method, edTPA panelists recommended an initial cut score, which was then discussed and evaluated. An additional round of recommendations was conducted during the session before the panel recommended a final cut score (edTPA Transition Plan, p.263-264).

[31] Appendices for this report are in Volume II: Appendices.

CalTPA Evidence Sources

Observation. Two HumRRO staff members observed CalTPA's Standard-Setting Panel Meeting in Sacramento, California on May 8 and 9, 2019. The purpose of the meeting was to convene an expert panel of educators to determine and recommend a passing standard for CalTPA. Like edTPA, facilitators used the briefing book method (Haertel, 2005, 2008).

Prior to the meeting, each panelist was asked to complete pre-work, which consisted of reviewing CalTPA Performance Assessment Guides (including rubrics), scoring materials, and six previously scored CalTPA submissions representing different performance levels across various content areas. The model sponsors provided panelists with instructions to assist their pre-work that included aspects of the materials on which to focus and two framing questions (see Appendix 5.B. for a copy of the instructions). The framing question for the review of the Performance Assessment Guide and rubrics was: "*Given the scope and contents of this CalTPA Cycle (the required evidence and rubrics), think about a teacher candidate who is just at the level of knowledge and skills required to perform effectively the job of a new teacher in California public schools*." The framing question for the review of the CalTPA submissions was: "*Does this candidate meet your definition of "a teacher candidate who is just at the level of knowledge and skills required to perform effectively the job of a new teacher in California public schools? Why or why not?"*

On the first day of the Standard-Setting Panel Meeting, the meeting facilitators (CalTPA representatives) provided an overview of the (a) panel's charge, (b) CalTPA's design, and (c) CalTPA's Scorer Training and Calibration. Then, facilitators conducted the workshop's "Standard Setting Policy Capture Jigsaw Activity." For this activity, the facilitators rotated panelists into four small groups (of 4-5 panelists) to review candidate portfolio submissions for six rounds (with 3 rounds for each of the assessment cycles). For both Cycle 1 and Cycle 2, eight of the 12 small groups reviewed a unique submission and four of small groups reviewed the same submission (because eight of the nine pre-work submissions were assigned to only four to six panelists and one was reviewed by all 21 panelists).

Within each small group of the "Standard Setting Policy Capture Jigsaw Activity," panelists first made individual ratings (on a scale that included "Clearly below," "Just below," "Just meets," and Clearly meets") for about 12 minutes then discussed their individual ratings and came to a group consensus rating for about 13 minutes (see Appendix 5.B. for a copy of the activity's instructions). After the Jigsaw activity was complete for each assessment cycle, Table Leaders (i.e., each of the small group leaders) presented rating results and rationales to the full panel on the candidate submissions. Based on these results and rationales, CalTPA staff facilitated a discussion intended to narrow the range of scores under consideration for the passing threshold for Cycle 1 and for Cycle 2. CalTPA facilitators collected all individual and group rating forms. After the session was complete, meeting facilitators compared the individual and small group ratings against the previously scored rating for each submission that was considered "correct." Ratings by the panelists largely matched the previously scored ratings.

On the second day of the Standard-Setting Panel Meeting, the facilitators presented a recap of the range of scores generated from the Day 1 activities. The facilitators reengaged the panelists in discussion on which of the reviewed submissions were within the passing range with the goal of narrowing the range of cut scores under consideration. They did this for both Cycle 1 and Cycle 2. Next, the facilitators presented a sample of Assessment Results Reports with rubric scores and a total cycle score for a sample of hypothetical teacher candidates. These reports, or score profiles, represented the range of scores that roughly corresponded to the range of

scores being considered for the cut score so that panelists could consider compensatory scoring and have concrete examples of a range of diverse candidate score profiles. Then, the facilitators presented the CalTPA descriptive statistics for the first operational year. Panelists reviewed the number of submissions by content area and the descriptive statistics (e.g., mean, median) overall and by rubric, content area, and demographic groups. Based on this information, the panelists were asked to independently make an initial cut score recommendation for Cycle 1. This information was captured on individual rating forms. This same process was repeated for Cycle 2. The facilitators then compiled the individual initial cut score recommendations for Cycle 1 and Cycle 2 and presented those results to the panelists. They engaged panelists in a discussion of the most frequently recommended cut score (i.e., mode cut score), the mean cut score recommendation, and the median cut score recommendation. Following the discussion of their initial cut score recommendations, the panelists were presented with impact data—that is, the percentage of candidates in the first operational year that would pass the CalTPA if the passing threshold was set at each score point being considered. Following discussion of the impact data, the facilitators asked panelists to provide a final independent recommendation for a passing standard for Cycle 1 and Cycle 2. The median panelist cut score for each cycle (a 19 for Cycle 1 and a 21 for Cycle 2) was presented to the whole group. The panelists also recommended that candidates must obtain at least a 2 on all rubrics in order to pass; that is, if a candidate obtains a 1 on any rubric, then they would not pass. This condition was included with the median cut score recommendations presented to the Commission.

Documentation. Review of materials and documentation provided to panelists for pre-work and at the Standard-Setting Panel Meeting complimented our May 2019 observation of CalTPA's standard setting process. We reviewed (a) CalTPA's Performance Assessment Guides and rubrics (b) all teacher candidate submissions (nine per cycle) provided to panelists, and (c) the Commission agenda item regarding the standard setting process that included a brief report of the task. The briefing book, a binder of paper materials provided to panelists and observers at the meeting, included the following eight tabs of information.

- Tab 1. CalTPA Design Team, CalTPA Key Milestones, Crosswalk Summary Chart – CalTPA and TPEs
- Tab 2. CalTPA Performance Assessment Guides for Multiple Subject Cycles 1 and 2
- Tab 3. CalTPA Performance Assessment Guides for Single Subject Cycles 1 and 2
- Tab 4: CalTPA Pre-Work for Panelists – Instruction and Process
- Tab 5: CalTPA Standard Setting Policy Capture Activity Instructions
- Tab 6: CalTPA Candidate Score Profiles
- Tab 7: CalTPA Standard Setting – Samples and Descriptives
- Tab 8: CalTPA Standard Setting – Impact Data

Evaluation Steps

To evaluate standard-setting procedures, we considered the information obtained from observation of standard-setting (where applicable) and standard-setting documentation for adherence to (a) the relevant *Assessment Design Standards* (ADS) and (b) the Standards relevant to standard-setting from the *Joint Standard*s (JS).[32] We identified four Standards

[32] We capitalize "Standard" throughout this chapter when referring to a standard specified by the ADS or the *Joint Standards*.

directly relevant to this activity. One Standard is ADS (1m); the other three Standards are from the *Joint Standards*.

To evaluate the extent to which the available information adheres to these Standards, two HumRRO researchers independently assigned a strength of evidence rating based on evidence reviewed using the rating scale presented in Table 5.1. The raters discussed and came to consensus on any discrepant ratings.

Table 5.1. Rating Scale for Strength of Evidence

Rating Level	Description of Rating Levels
1	No evidence of the Standard/element found in the documentation provided.
2	Little evidence of the Standard/element found in the documentation; less than half of the Standard/element covered in the documentation and/or evidence of key aspects of the Standard/element could not be found.
3	Some evidence of the Standard/element found in the documentation; approximately half of the Standard/element covered in the documentation including some key aspects of the Standard/element.
4	Evidence in the documentation mostly covers the Standard/element; more than half of the Standard/element covered in the documentation, including key aspects of the Standard/element.
5	Evidence in the documentation fully covers all aspects of the Standard/element.

Next, we developed a checklist to evaluate the details of the standard-setting process; the checklist was developed based on the criteria set forth by Hambleton (2001). These criteria address the features of a standard-setting, such as developing performance level descriptors; outlining the KSAs of a minimally competent examinee; training the participants; and the general standard-setting procedure qualities. If the identified feature was observed, a "√" was entered in the observation column. If the information was not available, then a rating of "CR" was assigned for cannot rate at this time. If a feature was not applicable to the specific standard-setting method or to a specific situation an "NA" was entered.

Results

We present the results of the numeric ratings assigned to each of the Standards using the rating scale presented in Table 5.1. The results of the ratings of the Standards are presented in Tables 5.2, 5.4, and 5.6 for FAST, edTPA, and CalTPA, respectively. Each table includes the (a) ADS and *Joint Standards* in the left column, (b) rating on the strength of evidence for the Standard in the middle column, and (c) rationale for the rating in the right column. Tables 5.3, 5.5, and 5.7 present the respective standard-setting process checklists for FAST, edTPA, and CalTPA.

FAST

Table 5.2 presents the ratings for FAST on each relevant ADS and *Joint Standard*. Note that both observations and supporting documentation were used to make the ratings presented in Table 5.2. Table 5.3 presents the FAST standard-setting process checklist.

Table 5.2. Ratings on the Assessment Design Standard and Joint Standards for FAST

Standards	FAST Rating	Rationale for FAST Rating
ADS 1(m): In the course of determining a passing standard, the model sponsor secures and reflects on the considered judgments of teachers, supervisors of teachers, 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 model sponsor periodically reviews the reasonableness of the scoring scales and established passing standard, when and as directed by the Commission.	4	The August 2017 FAST Preliminary Passing Standard Workshop included a review of the SVP and TPEs and a discussion of the rubrics by 12 participants. Five were master teachers, six were university coaches, and one was a Special Education faculty member. The group had a median of 1.5 years' experience with a variety of credential areas. Five were men and seven were women. Seven representatives were white, two were Asian, and three were Latino. During the SVP Passing Standard Workshop in May 2018, the model sponsor secured and reflected on the judgments of eight educator preparation experts with regard to the clarity and appropriateness of the SVP prompts and the Level 1 (Does Not Meet Expectations) and Level 2 (Meets Expectations) descriptors for the rubrics. All participants were white females. Four represented a single subject credential area and four represented the multiple subject credential area. The May 2018 TSP Passing Standard Workshop included six participants. All were university coaches. Half represented a Single Subject credential area and half represented the Multiple Subject credential area. All participants were white. One was male, and the rest were female. FAST could consider expanding its expert group that reviewed the rubrics to include other support providers of new teachers and attempt to compose a more diverse set of participants. The SVP and TSP events should have included participants who were not active university coaches.
JS 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.	2	The FAST model did not use a typical standard setting process (e.g. Body of Work, Bookmark) to determine the cut score. The FAST model should document its rationale for using a non-traditional standard setting method and include justification for why their non-traditional approach is appropriate for their needs.
JS 5.22: When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgements can bring their knowledge and experience to bear in a reasonable way.	4	During the SVP Passing Standard Workshop, participants were observed bringing their knowledge and experience to bear on the edits to the prompts and rubrics, but not on performance data (i.e., impact data) or actual candidate submissions.
JS 5.23: When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.	3	To aid in their discussion of the clarity and appropriateness of the rubric descriptors, participants referenced candidate performances that they had observed (i.e., from their memory). However, the participants did not have actual candidate submissions available during this discussion. Using actual candidate submissions to aid this discussion would further bolster support for this Standard.

Table 5.3. FAST Standard Setting Process Criteria Checklist

Checklist Items:	Information Source:		Description
	Documentation	Observation	
Was consideration given to the groups who should be represented on the standard-setting panel and the proportion of the panel that each group should represent?	√	√	Participants familiar with FAST were selected. Background and demographic information on the panelists were provided. The August 2017 FAST Preliminary Passing Standard Workshop included a review of the SVP and TPEs and a discussion of the rubrics by 12 participants. Five were master teachers, six were university coaches, and one was a Special Education faculty member. The group had a median of 1.5 years' experience with a variety of credential areas. Five were men and seven were women. Seven representatives were white, two were Asian, and three were Latino. During the SVP Passing Standard Workshop in May 2018, the model sponsor secured and reflected on the judgments of eight educator preparation experts with regard to the clarity and appropriateness of the SVP prompts and the Level 1 (Does Not Meet Expectations) and Level 2 (Meets Expectations) descriptors for the rubrics. All participants were white females. Four represented a single subject credential area and four represented the multiple subject credential area. The May 2018 TSP Passing Standard Workshop included six participants. All were university coaches. Half represented a Single Subject credential area and half represented the Multiple Subject credential area. All participants were white. One was male, and the rest were female.
Was the panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on the assessment?	√	√	The August 2017 FAST Preliminary Passing Standard Workshop included 12 participants, the May 2018 SVP Passing Standard Workshop included eight participants, and the May 2018 TSP Passing Standard Workshop included six participants. Overall, the August 2017 FAST Preliminary Passing Standard Workshop and May 2018 SVP Passing Standard Workshop panels appeared to be large enough for the purpose of this workshop in conjunction with the size and scope of FAST, a site-specific assessment model. However, the May 2018 TSP Passing Standard Workshop appeared to be small for its purpose.

(continued)

Table 5.3. (Continued)

Checklist Items:	Information Source:		Description
	Documentation	Observation	
Were two panels used to check the generalizability of the performance standards?		√	Subpanels were not formed.
Were subpanels within a panel formed to check the consistency of performance standards over independent groups.		√	Subpanels were not formed.
Was the performance standard-setting method field tested in preparation for its use in the standard-setting study, and revised accordingly?			No. This performance standard-setting method was not field tested in preparation for its use in the standard-setting study.
Is there documentation or discussion to suggest the selected method was tried out or was used in previous years?			No, there is no documentation to suggest the selected method was tried out or was used in previous years
Is there any indication that a technical advisory committee reviewed the standard-setting plan?		√	The standard-setting plan was not reviewed by a TAC.
Is there any indication that the facilitators had a pre-conceived idea of what the passing cut score should be? Describe.		√	The passing standard was previously set (i.e., a 2 on all rubrics); the discussion during the workshop did not concern changing the passing score, rather they verified that the language for the Level 2 (Meets Expectations) descriptor accurately described a just sufficiently qualified candidate.
Was the standard-setting method appropriate for the particular educational assessment and was it described in detail?			No documentation was available to describe how the cut score was set at 2, nor the rationale for the standard setting method that was used. The model sponsor should provide such documentation.
Was the purpose of the educational assessment and the uses of the test scores explained to panelists at the beginning of the standard seeing meeting?		√	Yes, the purpose of assessment and the use of scores was described to the panelists. The panelists were also familiar with the assessment due to their roles as university coaches and FAST scorers.
Were panelists exposed to the assessment itself and how it was scored?		√	Panelists were familiar with the assessment as university coaches and FAST scorers. An overview of the assessment, the TPEs, and the rubrics was also provided.

(continued)

Table 5.3. (Continued)

Checklist Items:	Information Source:		Description
	Documentation	Observation	
Were the auditors (HumRRO) provided materials to review in advance or at the workshop? Describe what was provided and if it was helpful.	√		The following documentation was provided: 1. Passing Standard Workshop Agenda SVP 2. Class Profile SVP FAST2.0 3. Lesson Plan Template 4. SVP 2.0 Fall 5. SVP rubric passing standard 6. Activity or Strategy Table SVP 2.0 7. FAST Passing Standard process See Appendix 5.A for additional description of these documents. The documents were helpful in understanding how the workshop was conducted.
What materials were provided to the panelists?	√	√	Materials 1-7 listed above were provided to the panelists during the meeting.
Were the rubrics/scoring rules described and understood by panelists?		√	Yes, the rubrics and levels were described and understood by the panelists.
Were the qualifications and other relevant demographic data about the panelists collected?	√		Yes, FAST facilitators collected participant demographics and background information.
Were panelists administered the educational assessment, or at least a portion of it?		√	The participants were not administered the assessment and did not view recordings of candidates performing the task. However, as university coaches/FAST scorers they were familiar with candidate performance on FAST.
Were panelists suitably trained on the method to set performance standards?			The workshop was a review of the SVP rubrics.
Describe training. Was there pre-work? Were there checks of understanding?			There was no pre-work. Panelists participated in a brainstorming exercise to outline skills essential for beginning teachers. They also held a discussion on the appropriateness and reasonableness of the KSAs associated with rubric Level 2 (Meets Expectations) to classify a minimally qualified teacher.
Were descriptions of the performance categories sufficiently clear so as to allow panelists to accurately apply the categories within standard-setting process?		√	The panelists were tasked with determining if the rubrics were appropriate and reflected reasonable expectations of the KSAs for beginning teachers.

(continued)

Table 5.3. (Continued)

Checklist Items:	Information Source:		Description
	Documentation	Observation	
Were just minimally qualified candidate Performance Level Descriptors (PLDs) developed for each cut? How were they described?		√	No minimally qualified candidate PLDs were developed for this workshop.
If an iterative process was used for discussing and reconciling rating differences, was feedback to panelists clear, understandable, and useful?		√	There were no ratings collected from the panelists. Panelists provided qualitative input on the clarity and reasonableness of the rubric, focusing on the Level 2 (Meets Expectations) rubric.
If the process was iterative, were there any unusually large changes in the cut score? If so, why?		√	NA
Were panelists' ratings captured on paper or via computer?		√	NA
Were the rating forms easy to use?		√	NA
Did panelists have computer or technical issues when making their ratings?		√	NA
Were documents such as candidate booklets, tasks, items, and so on, simply coded?		√	No candidate data was used in this workshop
Was the process conducted efficiently?		√	The participants worked together effectively to reach consensus regarding their opinion of the appropriateness of the rubric. The process was conducted efficiently; the facilitator kept the discussion focused and was able to prompt the panelists and summarize their decisions.
Were panelists given the opportunity to “ground” their ratings with performance data and how was the data used?		√	No performance data was used in this workshop.
Were panelists provided consequential data (or impact data) to use in their deliberations and how did they use the information?		√	No impact data was used in this workshop.
Was the approach for arriving at the final performance standards clearly described and appropriate?		√	The approach for tweaking the rubric descriptors was clear and appropriate.
Was a final evaluation of the process conducted?		√	No, a final evaluation of the process was not conducted.

(continued)

Table 5.3. (Continued)

Checklist Items:	Information Source:		Description
	Documentation	Observation	
Was evidence compiled to support the validity of the performance standards?		√	No
Was the full standard-setting process documented (from the early discussions of the composition of the panel to the compilation of the validity evidence to support the performance standards)?	√	√	Documentation of the full standard-setting process was not available.
Were effective steps taken to communicate the performance standards to others after standard-setting?		√	The intent is to post the rubric on the Commission web site when it is finalized.
Is HumRRO able to obtain results of administration?	√		Score data was provided via a request for Activity 6 (see Chapter 6)

edTPA

Table 5.4 presents the ratings for edTPA on each ADS and *Joint Standard*. Note that we were only able to use supporting documentation to make the ratings on the Standards and standard-setting checklist given that the standard-setting workshop occurred July 1, 2014 (i.e., prior to the start of the comparability study). Table 5.5 presents the edTPA standard-setting process checklist.

Table 5.4. Ratings on the Assessment Design Standard and Joint Standards for edTPA

Standards	edTPA Rating	Rationale for edTPA Rating
ADS 1(m): In the course of determining a passing standard, the model sponsor secures and reflects on the considered judgments of teachers, supervisors of teachers, 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 model sponsor periodically reviews the reasonableness of the scoring scales and established passing standard, when and as directed by the Commission.	5	The model sponsor used the "Briefing Book" method for determining a passing standard, which included consideration of judgements of active California faculty members from institutions of higher education, TK–12 educators, and members from various California stakeholder groups. An overview of the Briefing Book Standard Setting Method was provided in the Transition Plan (see p.263). A detailed introduction of the standard-setting process was also provided in the Transition Plan (p.445-448). In addition, as indicated in the annual administrative reports, the standard-setting process has been repeated in prior years and the reasonableness of the scoring scales has been reviewed.
JS 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.	5	The rationale and procedures used for establishing cut scores were documented clearly by edTPA in an overview of the Briefing Book Standard-Setting Method and detailed standard-setting process description provided in the Transition Plan (see p.263 and 445-448).

(continued)

Table 5.4 (Continued)

Standards	edTPA Rating	Rationale for edTPA Rating
JS 5.22: When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgements can bring their knowledge and experience to bear in a reasonable way.	5	Panelists were informed of the purpose of the assessment and are provided with the "Briefing Book" to guide their activity. Prior to the meeting, each invited panelist received edTPA handbooks, rubrics, scoring materials, and three previously scored sample portfolio submissions representing different performance levels across various content areas. Panelists were asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions assigned to them. During the facilitated session, panelists familiarized themselves with the assessment and with the information contained in the briefing book. After a series of "Policy Capture Activities" examining whole portfolios and score profiles representing a range of candidate performances, panelists recommended an initial cut score (also be referred to as a "passing standard") for each task, which was then discussed and evaluated based on impact data. Following that, panelists recommended a final cut score (Transition Plan, p.307).
JS 5.23: When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.	5	The evidence-based process was informed by state-specific data as well as comparative (non-California) data (p. 69 of the edTPA Transition Plan).

Table 5.5. edTPA Standard Setting Process Criteria Checklist

Checklist Items:	Information Source		Description
	Documentation	Observation	
Was consideration given to the groups who should be represented on the standard-setting panel and the proportion of the panel that each group should represent?	√		The California standard setting included 12 panelists: Seven panelists from institutions of higher learning, one active full-time teacher, one panelist who was both a high school teacher and worked for a college, one panelist who worked for the California Council of Teacher Education and a university, one panelist who worked for the California Teachers Association and a university, and one panelist who worked for a county office of education. Consideration was not given to demographic characteristics of the participants. However, their experience was taken into consideration. Length of experience was also not described.

(continued)

Table 5.5. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Was the panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on the assessment?	√		The California standard setting consisted of 12 participants who represented institutions of higher learning, various associations, and teachers. The panel's representativeness of various demographic groups and other constituencies is unknown. Given the size and number of constituencies this panel represented, it was small.
Were two panels used to check the generalizability of the performance standards?	√		The California standard setting did not use multiple panels.
Were subpanels within a panel formed to check the consistency of performance standards over independent groups.	√		See above.
Was the performance standard-setting method field tested in preparation for its use in the standard-setting study, and revised accordingly?			Field testing of the standard-setting was not described in the report, but it is reasonable to assume that this method was field tested previously based on the overview of the standard-setting method (p.263).
Is there documentation or discussion to suggest the selected method was tried out or was used in previous years?			Documentation or discussion to suggest the selected method was tried out or was used in previous years is not described in the report, but it is reasonable to assume that this method was field tested previously based on the overview of the standard-setting method (p.263).
Is there any indication that a technical advisory committee reviewed the standard-setting plan?			There is no indication in the Transition Plan report that a technical advisory committee reviewed the standard-setting plan.
Is there any indication that the facilitators had a pre-conceived idea of what the passing cut score should be? Describe.			There is no indication in the Transition Plan report that facilitators had a pre-conceived idea of what the passing cut score should be.
Was the standard-setting method appropriate for the particular educational assessment and was it described in detail?	√		Yes, detailed description is provided on page 263 of Transition Plan.
Was the purpose of the educational assessment and the uses of the test scores explained to panelists at the beginning of the standard-setting meeting?	√		Yes, the panelists reviewed the assessment for homework and thus were familiar with the assessment, its purposes, and scoring system.

(continued)

Table 5.5. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were the auditors (HumRRO) provided materials to review in advance or at the workshop? Describe what was provided and if it was helpful.	√		The standard-setting was conducted in 2014; however, the HumRRO reviewers were able to access the standard-setting documentation from the edTPA Transition Plan.
What materials were provided to the panelists?	√		Prior to the standard setting, each invited panelist received edTPA handbooks, rubrics, scoring materials, and three previously scored edTPA submissions representing different performance levels across various content areas. Panelists were asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions that were assigned to them (Transition Plan, p.446).
Were the rubrics/scoring rules described and understood by panelists?	√		Yes, the rubrics and scoring rules were described to the panelists. However, the report does not contain reference to checks for understanding of instructions by the panelists.
Were the qualifications and other relevant demographic data about the panelists collected?	√		The panelists' place of employment information was collected; however, demographic information and years of experience were not available in the report.
Were panelists administered the educational assessment, or at least a portion of it?	√		Prior to the meeting, each invited panelist received edTPA handbooks, rubrics, scoring materials, and three previously scored edTPA submissions representing different performance levels across various content areas. Panelists were asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions that were assigned to them. (Transition Plan, p.446).
Were panelists suitably trained on the method to set performance standards?	√		Yes, according to the description of standard-setting panelists received training regarding purposes and methods of standard-setting.

(continued)

Table 5.5. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Describe training. Was there pre-work? Were there checks of understanding?	√		Prior to the meeting, each invited panelist received edTPA handbooks, rubrics, scoring materials, and three previously scored edTPA submissions representing different performance levels across various content areas. Panelists were asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions that were assigned to them. Throughout the standard-setting event and examination of sample edTPA submissions, a guiding question was used and revisited to frame all discussions, which provided a common ground from which all participants could anchor their passing standard judgements. The guiding question was this: *What score (the sum of all the rubric scores of edTPA) represents the level of performance that would be achieved by a teacher candidate who is just at the level of knowledge and skills required to perform effectively the job of a new teacher?* The report does not describe checks for understanding.
Were descriptions of the performance categories sufficiently clear to allow panelists to accurately apply the categories within standard-setting process?	√		The report does not explicitly describe the KSAs associated with each performance level. It does say that the participants received the performance level of each assigned submission for every rubric. They were then asked to rate each of those submissions as Clearly Below, Just Below, Just Meets, or Meets the Standard. No operationalizations (i.e., definitions) of these demarcations were provided.

(continued)

Table 5.5. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were just minimally qualified candidate PLDs developed for each cut? How were they described?	√		PLDs were not developed. Panelists rated portfolios as "Clearly Below," "Just Below," "Just Meets," or "Meets the Standard." Throughout the standard-setting event and examination of sample edTPA submissions, a guiding question was used and revisited to frame all discussions, which provided a common ground from which participants could anchor their individual judgements of an appropriate passing standard. • Think about a teacher candidate who is just at the level of knowledge and skills required to perform effectively the job of a new teacher in California public schools. • Guiding question: What score (the sum of all of the rubric scores of the edTPA) represents the level of performance that would be achieved by this individual?
If an iterative process was used for discussing and reconciling rating differences, was feedback to panelists clear, understandable, and useful?	√		To begin, each panelist spent some time recalling a specific submission that they reviewed for homework and then provided an individual rating for that portfolio. Panelists rated portfolios as Clearly Below, Just Below, Just Meets, or Meets the Standard. Then, in assigned table groups, panelists discussed their ratings with other panelists with the goal of arriving at a consensus rating. Upon reaching consensus, each table completed one consensus rating form for the portfolio discussed. After each table completed the table form, panelists moved to the next table assignment and they repeated the process two more times for the other submissions they reviewed for homework. By the end of the three cycles, a consensus rating was generated for each of the submissions reviewed by each panel (Transition Plan, p.307-308). After the series of activities examining whole portfolios and score profiles representing a range of candidate performances, panelists recommended an initial cut score (which may also be referred to as a "passing standard") for each task, which was then discussed and evaluated based on impact data. Following that, panelists recommended a final cut score (Transition Plan, p.382).

(continued)

Table 5.5. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
If the process was iterative, were there any unusually large changes in the cut score? If so, why?			It is unknown if unusually large changes in the cut-score occurred during the process. The ratings and differences between ratings are not mentioned in the report.
Were panelists' ratings captured on paper or via computer?			Unknown; this information is not available, because we did not observe the standard-setting and it was not mentioned in the report.
Were the rating forms easy to use?			Unknown; this information is not available, because we did not observe the standard-setting.
Did panelists have computer or technical issues when making their ratings?			Unknown; this information is not available, because we did not observe the standard-setting.
Were documents such as candidate booklets, tasks, items, and so on, simply coded?			Unknown; this information is not available, because we did not observe the standard-setting.
Was the process conducted efficiently?	√		It can be inferred from the report that the process was conducted efficiently because it was well planned, and the facilitators had experience conducting similar standard-setting studies.
Were panelists given the opportunity to "ground" their ratings with performance data and how was the data used?	√		Panelists were provided descriptive and summary data to help guide their recommendations. Descriptive and summary data included the number of portfolios scored in each edTPA credential field, a summary of the population aggregate rubric, task, and total edTPA performance (mean, standard deviation, median, minimum, maximum) for all candidates. Demographics and total score descriptive performance statistics (number, percent, mean, standard deviation, and median, minimum, maximum) were provided by gender, ethnicity, and Primary Language English subgroups. Finally, a distribution of total scores was provided from the national data set (Transition Plan, p.308).

(continued)

Table 5.5. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were panelists provided consequential data (or impact data) to use in their deliberations and how did they use the information?	√		Panelists were provided impact data to help guide their recommendations. Impact data included the reporting of the passing rate that would have been observed based on the range of possible cut scores determined in Policy Capture 1. Included in the impact data were comparisons between the host state (i.e. California) and other states where edTPA is non-consequential. The number of candidates passing and the passing rate (as a percentage of all candidates in a given group) overall, by credential area, and by demographic characteristics were also provided (Transition Plan, p.308).
Was the approach for arriving at the final performance standards clearly described and appropriate?	√		After a series of activities examining whole portfolios and score profiles representing a range of candidate performances, panelists recommended an initial cut score (which may also be referred to as a "passing standard") for each task, which was then discussed and evaluated based on impact data. Following that, panelists recommended a final cut score (Transition Plan, p.382). This process was appropriate based on the structure of this standard-setting method.
Was a final evaluation of the process conducted?			It is unknown if a final evaluation of the process was conducted. his information was not available from the report.
Was evidence compiled to support the validity of the performance standards?	√		Included in the impact data were comparisons between the host state (i.e. California) and other states where edTPA is non-consequential. The number of candidates passing and the passing rate (as a percentage of all candidates in a given group) overall, by credential area, and by demographic characteristics were also provided (Transition Plan, p.308). Other validity studies that related performance to other criteria (e.g., other assessments) was not included.
Was the full standard-setting process documented (from the early discussions of the composition of the panel to the compilation of the validity evidence to support the performance standards)?	√		The standard-setting process was documented in standard-setting reports and in the Transition Plan.

(continued)

Table 5.5. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were effective steps taken to communicate the performance standards to others after standard-setting?	√		Documentation of standard-setting was communicated to audiences via the standard-setting reports and the Transition Plan.
Is HumRRO able to obtain results of administration?	√		Yes, score data was provided via a request for Activity 6 (see Chapter 6)

CalTPA

Table 5.6 presents the ratings for CalTPA on each ADS and *Joint Standard*. Table 5.7 presents the CalTPA standard-setting process checklist.

Table 5.6. Ratings on the Assessment Design Standard and Joint Standards for CalTPA

Standards	CalTPA Rating	Rationale for CalTPA Rating
ADS 1(m): In the course of determining a passing standard, the model sponsor secures and reflects on the considered judgments of teachers, supervisors of teachers, 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 model sponsor periodically reviews the reasonableness of the scoring scales and established passing standard, when and as directed by the Commission.	5	CalTPA convened twenty-one content experts in May 2019 to recommend the passing standard based on discussion of necessary and acceptable levels of proficiency on the part of entry-level teachers.[33] The "Briefing Book" method was used. Panel members included teachers and TK-12 district-level staff (n = 3), associations (n = 1), university faculty (n = 16), and one retired educator. Eleven panelists had been lead assessors, seven had been regular assessors, and five participated in the design team for the CalTPA. The panel was composed largely of teacher preparation program faculty and staff. There were few panelists who directly provide support to new teachers who have licensure. It is unknown how often the passing standard will be reviewed, however, near the end of the standard-setting meeting, Amy Reising, CalTPA's Director of Performance Assessment Development said that the cut-scores would likely be revisited every 2-3 years.

(continued)

[33] Note that one panelist was unable to attend the second day of the standard setting.

Table 5.6. (Continued)

Standards	CalTPA Rating	Rationale for CalTPA Rating
JS 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.	5	CalTPA reported the rationale, procedures, and results of the standard-setting meeting to the Commission and posted it online. The briefing book materials used at the meeting included an overview of the method itself and detailed instructions that included clearly documented procedures for individual and small group consensus ratings.
JS 5.22: When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgements can bring their knowledge and experience to bear in a reasonable way.	5	During the standard setting, panelists took part in a "Jigsaw" style activity for each cycle to discuss candidate submissions and make ratings of "Clearly below," "Just below," "Just meets," and "Clearly meets." In conducting these activities, different compositions of small groups of panelists reviewed different candidate submissions (each panelist reviewed three submissions per cycle). One candidate submission per cycle was reviewed by all panelists. Overall, the method provided panelists with the opportunity to voice their judgement and hear other perspectives. At the end of the standard setting, initial and final recommendations were made by tallying individual ratings rather than a consensus of panelists ratings so that each panelist's judgements were weighted equally.
JS 5.23: When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.	5	CalTPA facilitators provided panelists with access to Cycle 1 and Cycle 2 first year operational performance data (i.e., impact data), which allowed panelists to determine the percentage of candidates who would pass at each cut-score being considered. The 2018–19 data were pulled as late as possible (on April 25, 2019 when there were 6,386 submissions) to ensure as large and representative of a sample as possible. Relationships of performance data to other criteria (e.g., concurrent validity studies) were not included.

Table 5.7. CalTPA Standard Setting Process Criteria Checklist

Checklist Items:	Information Source		Description
	Documentation	Observation	
Was consideration given to the groups who should be represented on the standard-setting panel and the proportion of the panel that each group should represent?	√	√	Panel members included teachers and TK-12 district-level staff (n = 3), association staff members (n = 1), university faculty/staff (n = 16), and a retired educator (n = 1). A concerted effort was made to ensure most panelists had familiarity with the assessment, its design, and its rubrics. Eleven panelists had been lead assessors, seven had been regular assessors, and five participated in the CalTPA design team. A few panelists had been working on CalTPA for 20 years, since its inception. Panelists applied to participate in the standard-setting meeting by completing a form that asked for 1) content area expertise, 2) test development and scoring experience, 3) TK-12 teaching experience including geographic location, and ethnicities taught, 4) college/university experience including institution and exposure to requirements of teacher candidates related to CalTPA, 5) California teaching or administrative credential held, 6) highest level of education attained, 7) National Board of Teaching Certification, 8) world language fluency, 9) affiliated professional organizations, 10) race/ethnicity, and 11) gender. The data from the application form was not provided to HumRRO and thus the counts and percentages of panelists with these background/demographic characteristics is unknown. Content expertise was varied but did not cover all CalTPA's 16 content areas.
Was the panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on the assessment?	√	√	The panel was large enough to be suitable for the task of standard-setting. CalTPA convened twenty-one content experts (with one needing to drop after day 1). Panel members included teachers and TK-12 district-level staff (n = 3), associations (n = 1) university faculty (n = 16), and one retired educator. Eleven panelists had been lead assessors, seven had been regular assessors, and five participated in the CalTPA design team. Only two or three panelists were current TK-12 teachers. The panel was composed largely of teacher preparation program faculty and staff. There were few panelists who directly provide support to new teachers who have licensure. Panelists represented a diverse array of backgrounds and perspectives. Counts and percentages of all applicable demographic and background characteristics is unknown.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were two panels used to check the generalizability of the performance standards?	√	√	One large, 21-person panel was formed to conduct the standard-setting. During the standard setting, panelists took part in "jigsaw" style activities for each cycle to discuss candidate submissions, make ratings, and narrow the range of scores under consideration for the passing threshold. In conducting these activities, different compositions of small groups of panelists reviewed different candidate submissions (each panelist reviewed three submissions per cycle). One candidate submission per cycle was reviewed by all panelists. Overall, the method provided panelists with exposure to different perspectives and allowed for each panelist's opinions to be heard. Panelist discussion was the primary component of activities related to rubric examination and impact data. At the end of the standard setting, initial and final recommendations were made by tallying individual ratings rather than a consensus of panelists ratings so that each panelist's judgements were weighted equally. The 21-person panel was not split into two separate panels. Thus, there was no cross-check on the generalizability of the performance cut across split panels.
Were subpanels within a panel formed to check the consistency of performance standards over independent groups.	√	√	No. See above.
Was the performance standard-setting method field tested in preparation for its use in the standard-setting study, and revised accordingly?	√	√	The exact procedures for this standard-setting were not field tested. However, CalTPA used the "Briefing Book" standard-setting method, which is a method program staff have utilized for other assessments (including edTPA) in the past. The method was presented to the Commission prior to being utilized.
Is there documentation or discussion to suggest the selected method was tried out or was used in previous years?	√	√	The May 2019 standard-setting was the first time this activity took place for the newly updated CalTPA. Prior versions of the CalTPA used similar methods to set the passing standard. Of note, documentation of the "Briefing Book" method as selected and employed by edTPA is available.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Is there any indication that a technical advisory committee reviewed the standard-setting plan?	√		CalTPA communicated to the Commission that it was conducting a standard-setting meeting using the "Briefing Book" method. A technical advisory committee did not review the standard-setting plan.
Is there any indication that the facilitators had a pre-conceived idea of what the passing cut score should be? Describe.	√	√	The facilitators did not indicate their idea of a specific cut score and encouraged panelists to make their own judgements. For example, on Day 1, a panelist explained that he was coming to the meeting with a preconceived notion of what the cut-score should be and a CalTPA facilitator responded to the statement by suggesting the panelist ground his judgment in the evidence (example submissions) and refer to rubrics.
Was the standard-setting method appropriate for the particular educational assessment and was it described in detail?	√	√	The standard-setting method was appropriate for the CalTPA and its design. The briefing book method has been successfully used since the early 2000s for similar performance assessments (Haertel, 2008). CalTPA implemented it appropriately. Looking at actual candidate submissions was appropriate for setting a cut score on this assessment. Presentations and materials provided at the meeting described the standard setting procedures in detail. During the meeting, CalTPA utilized a "jigsaw activity" for panelists to review candidate submissions—not all panelists were given the same submission to review. There was only one common submission that all panelists reviewed for each Cycle. The upside of this approach was that 18 different submissions were reviewed by at least a handful of candidates. The downside of this approach was that there was only one submission that all the candidates saw, and each panelist only reviewed three submissions. The common submission (Multiple Subject – Mathematics) had a score of 18 for Cycle 1 and score of 20 for Cycle 2, which was within the range of the cut score being discussed. However, not all panelists saw common submissions around that score point when reviewing their other two submissions, which may have made it challenging for many panelists to make

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
(continued) Was the standard-setting method appropriate for the particular educational assessment and was it described in detail?	√	√	fine-grain distinctions between setting a cut at 18 vs 19, for example. Panelists did not define a Performance Level Description for the just sufficiently qualified candidate, aside from it being a mix of 2 and 3 ratings on the scoring rubric. While in small groups (i.e., jigsaw activity), panelists were overhead expressing varying definitions of "just meets the passing standard." One table defined "just meets the passing standard" as a teacher candidate who needs coaching, whereas they defined "clearly below the passing standard" as a teacher candidate who needs another class. Another table defined "just meets the passing standard" as a teacher candidate who would not do any harm to students, whereas they defined "clearly below the passing standard" as someone who may do harm to students. It would have been helpful to capture these discussions, bring them to the large group for discussion, and come to a common consensus understanding of how to operationalize the distinction between the various standard-setting policy capture categories of "Clearly below," "Just below," "Just meets," and "Clearly meets." Also, having all panelists see the same submissions around the cut point (i.e., 18, 19, 20 for Cycle 1 and 20, 21, and 22 for Cycle 2) and identifying a common operationalization of the categories may have helped to improve panelists' confidence in their cut score recommendations.
Was the purpose of the educational assessment and the uses of the test scores explained to panelists at the beginning of the standard-setting meeting?	√	√	Panelists reviewed the Performance Assessment Guides and rubrics for pre-work and meeting facilitators provided a presentation on CalTPA's assessment design, its purposes, and scoring system on Day 1 of the standard-setting meeting. On Day 2, the use of test scores was explained as one of multiple measures that factor into the credentialing decision. The subtleties of how CalTPA is used both for formative purposes and as a high-stakes assessment was not clear from the materials and presentations.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were panelists exposed to the assessment itself and how it was scored?	√	√	Yes, in addition to past experiences with CalTPA as local program coordinators and assessors, the panelists reviewed the Performance Assessment Guides and rubrics for pre-work and meeting facilitators provided a presentation on CalTPA's assessment design, its purposes, and scoring system. Additionally, panelists reviewed a sample of candidate submissions (three submission per Cycle).
Were the auditors (HumRRO) provided materials to review in advance or at the workshop? Describe what was provided and if it was helpful.	√	√	HumRRO observers were provided with the standard-setting agenda, current Performance Assessment Guides (including rubrics), pre-work instructions and process document, and all candidate submissions that were to be reviewed at the standard-setting meeting. These materials provided context and knowledge for the HumRRO observers to follow along with discussions by panelists and prepare for the meeting.
What materials were provided to the panelists?	√	√	Prior to the meeting, each invited panelist received CalTPA Performance Assessment Guides (including rubrics), pre-work instructions and process document, and six previously scored CalTPA submissions representing different performance levels across various content areas (three per Cycle). Panelists were asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions that were assigned to them. Panelists and auditors (HumRRO) were also provided a binder upon check-in. These binders included a Final Passing Standard form; standard setting meeting agenda; meeting evaluation form; overview of the briefing book method; professional norms for panelists; list of CalTPA Design Team members; CalTPA milestones (standard setting is the next to last milestone); crosswalk summary chart between CalTPA and TPEs; Performance Assessment Guides for MS Cycle 1 and Cycle 2: Performance Assessment Guides for SS Cycle 1 and Cycle 2; Copy of prework instructions for panelists; CalTPA Standard Setting Policy Capture Activity Instructions; CalTPA Candidate Score

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
(continued) What materials were provided to the panelists?	√	√	Profiles (sample candidate score reports across a range of scores for Cycle 1 and Cycle 2); CalTPA Standard Setting Samples and Descriptives; and Impact Data.
Were the rubrics/scoring rules described and understood by panelists?	√	√	Yes, the rubrics and scoring rules were provided to panelists with their pre-work materials and described to the panelists at the start of the first day of the standard-setting meeting. Because the panelists were invited to the meeting based on their experience and prior knowledge of the rubrics/scoring rules (as assessors and local CalTPA program coordinators), checks for understanding of the rubrics and scoring rules were not conducted. Questions about these materials were encouraged.
Were the qualifications and other relevant demographic data about the panelists collected?	√		As part of the standard-setting panelist application, the standard setting facilitators collected the panelists' place of employment information and other relevant experience related to CalTPA. It is unknown if demographic information was collected.
Were panelists administered the educational assessment, or at least a portion of it?	√	√	Panelists were not administered the assessment, which was appropriate given the time, effort, and access to students required to create a submission. Prior to the meeting, each invited panelist received CalTPA Performance Assessment Guides (including rubrics), scoring materials, and six previously scored CalTPA submissions representing different performance levels across various content areas. Panelists were asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions that were assigned to them.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were panelists suitably trained on the method to set performance standards?	√	√	The facilitators provided read ahead materials to panelists and presented training slides to panelists during the meeting. Facilitators allowed ample time for panelists to ask questions and for discussion. Having the framing question printed for each panelist to refer to throughout the meeting may have helped to clarify the objective—there were some panelists questions/comments that arose just prior to setting the final cut that suggested some lack of clarity on the objective (e.g., "We haven't discussed what percentage of candidates we think should pass" and "I think the cut score should be aspirational."). Facilitators responded appropriately, but it would have been helpful to have the specific, concrete framing question visible at all times.
Describe training. Was there pre-work? Were there checks of understanding?	√	√	Prior to the meeting, each invited panelist received CalTPA Performance Assessment Guides (including rubrics), scoring materials, and six previously scored candidate submissions representing different performance levels across various content areas. Panelists were asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions that were assigned to them. Some panelists said that they did not complete the prework. On Day 1 of the meeting, presentations about CalTPA's design and scoring were provided prior to a standard-setting policy "jig-saw" activity to discuss the candidate submissions provided for pre-work. On day 2, performance data and rubric language were reviewed, and discussions were held about panelists' rationales for setting cut-scores at various levels. Panelists were asked if they had questions. There were no formal checks of understanding.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Were descriptions of the performance categories sufficiently clear to allow panelists to accurately apply the categories within standard-setting process?	√	√	The "Clearly below the passing standard," "Just below the passing standard," "Just meets the passing standard," and "Clearly meets the passing standard" performance categories were somewhat vague; performance categories were not defined in terms of the KSAs required to be an effective beginning teacher. However, the requisite KSAs to be an effective beginning teacher are in the rubrics. Facilitators did repeatedly ask panelists to anchor their judgments in the rubrics.
Were just minimally qualified candidate PLDs developed for each cut? How were they described?	√	√	Panelists did not develop a PLD for the minimally qualified candidate. Essentially, panelists decided that a minimally qualified candidate exhibited KSAs that were defined by a combination of the Level 2 and Level 3 rubrics, but there was no synthesized PLD created for the minimally qualified candidate. The facilitators provided some guidance in the Policy Capture Activity Instructions (see Appendix 5.B. for a copy), which stated that *Just meets the passing standard = "JUST MEETS **your definition*** [emphasis added] *of performing effectively the job of a new teacher. This teacher has demonstrated some consistent strengths in teaching knowledge and skills and has a foundation on which to build. The teacher may have shown one or more minor flaws in teaching knowledge and skill that will likely improve with more time and experience."* It's likely that panelists may have differed with regard to "your definition of performing effectively the job of a new teacher," as there was no formal, facilitated discussion of how each panelist defined this.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
If an iterative process was used for discussing and reconciling rating differences, was feedback to panelists clear, understandable, and useful?	√	√	On Day 2 each panelist was asked to independently set a cut score for Cycle 1 and Cycle 2. Then, the facilitators showed the panelists the independently-derived cut scores (mean, median, and frequency distribution of cuts). The facilitators elicited discussion among panelists by asking to hear the rationale from the panelists who set the cut near the low end of the range of cuts, followed by asking to hear the rationale from the panelists who set the cut at the high end. When a panelist made a comment that was not aligned with the objective (e.g., "I don't want to fail someone because I don't want them to have to pay another $150 to take the test"), then the facilitators provided feedback in the form of re-orienting them to the objective of the standard setting in a clear, understandable, and useful way. Following the discussion of the Round 1 cuts, the panelists then made their final recommendations for the Cycle 1 and Cycle 2 cut-scores.
If the process was iterative, were there any unusually large changes in the cut score? If so, why?			The CalTPA standard-setting process was iterative but no unusually large changes in the cut score were made across those iterations. During day 1, the range of cut scores was narrowed to 4-5 cut scores per cycle. After CalTPA facilitators tallied the initial and final ratings on day 2, the average and median cut scores were within the narrowed range in both instances. The panelists' average and median initial cut scores and final cut scores were very close.
Were panelists' ratings captured on paper or via computer?		√	Paper
Were the rating forms easy to use?	√	√	The Initial Passing Standard Recommendation form and the Final Passing Standard Recommendation form were one page each, with clear instructions (see Appendix 5.C. for a copy of the final rating form).
Did panelists have computer or technical issues when making their ratings?			Not applicable. Computers were not used.
Were documents such as candidate booklets, tasks, items, and so on, simply coded?			The candidate submissions were each numbered and labeled.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Was the process conducted efficiently?	√	√	The Standard Setting was designed to be a 2-day meeting. The facilitators allowed time for panelists' questions and discussion, but kept the process moving forward so that it finished on time. If panelists had questions that weren't directly related to standard setting, then there was a designated process for addressing those (write down on a sticky note and post on a poster board labeled "parking lot"). This helped panelists feel that their questions were being heard/addressed, while keeping the process moving forward.
Were panelists given the opportunity to "ground" their ratings with performance data and how was the data used?	√	√	Meeting facilitators provided panelists with descriptive and summary data to help guide their recommendations. Descriptive and summary data included the number of submissions scored in each CalTPA content area, a summary of the population aggregate rubric, cycle, and total CalTPA performance (mean, standard deviation, median, minimum, maximum) for all candidates. Demographic and total score descriptive performance statistics (number, percent, mean, standard deviation, and median, minimum, maximum) were provided by gender, ethnicity, and language (e.g., English Only or multilingual), and setting subgroups. Additionally, each panelist reviewed six candidate submissions. Panelists were asked to refer to the submissions in helping them determine what constituted a minimally qualified teacher candidate.
Were panelists provided consequential data (or impact data) to use in their deliberations and how did they use the information?	√	√	Panelists were asked to make initial independent cut score determinations prior to seeing the impact data (Round 1). After making those initial cut score recommendations, impact data was shared with the panelists and they had an opportunity to change their initial cut score recommendation (Round 2). There was little change from the Round 1 to Round 2 cut-scores.

(continued)

Table 5.7. (Continued)

Checklist Items:	Information Source		Description
	Documentation	Observation	
Was the approach for arriving at the final performance standards clearly described and appropriate?	√	√	Panelists gained clarity on the approach for setting the cut as they went through the process. After a series of activities examining full candidate submissions and score profiles representing a range of candidate performances, panelists recommended an initial cut score (which may also be referred to as a "passing standard") for each cycle, which was then discussed and evaluated based on impact data. Following that, panelists recommend a final cut. This process was appropriate based on the structure of this standard-setting method.
Was a final evaluation of the process conducted?	√	√	Panelists were asked to complete an evaluation form prior to departing.
Was evidence compiled to support the validity of the performance standards?	√	√	Relationships of performance data to other criteria (e.g., concurrent validity studies) were not included.
Was the full standard-setting process documented (from the early discussions of the composition of the panel to the compilation of the validity evidence to support the performance standards)?	√		CalTPA reported the rationale, procedures, and results of the standard-setting meeting to the Commission and posted it online.
Were effective steps taken to communicate the performance standards to others after standard-setting?	√		CalTPA reported the rationale, procedures, and results of the standard-setting meeting to the Commission and posted it online. CalTPA used email to distribute the performance standards to teacher preparation program coordinators.
Is HumRRO able to obtain results of administration?	√	√	HumRRO will be provided with raw score data from the 2018–19 CalTPA administration in summer 2019.

Ratings Comparison

A comparison of ratings across models on the ADS and *Joint Standards* is presented in Table 5.8.

Table 5.8. Comparison of Ratings on Assessment Design/Joint Standard across TPAs

Standards	FAST Rating	edTPA Rating	CalTPA Rating
ADS 1(m) In the course of determining a passing standard, the model sponsor secures and reflects on the considered judgments of teachers, supervisors of teachers, 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 model sponsor periodically reviews the reasonableness of the scoring scales and established passing standard, when and as directed by the Commission.	4	5	5
JS 5.21 When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.	2	5	5
JS 5.22 When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgements can bring their knowledge and experience to bear in a reasonable way.	4	5	5
JS 5.23 When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.	3	5	5
Average	3.25	5.00	5.00

Discussion

The purpose of this activity was to investigate Claim 5: "*The standard-setting procedures used for each TPA model are sufficiently comparable and rigorous to ensure that the respective passing standards for each model accurately and consistently identify candidates possessing the requisite KSAs required to effectively teach the content area(s) authorized by the credential*."

After review of the evidence for all three TPA models related to standard-setting, it appears that edTPA and CalTPA use procedures that are sufficiently comparable and rigorous to ensure that their passing standards accurately and consistently identify candidates possessing the requisite KSAs required to effectively teach the content area(s) authorized by the credential. Across ADS 1(m) and *Joint Standards* 5.21, 5.22, and 5.23, HumRRO staff rated edTPA and CalTPA a "5" on every standard, which indicated that the evidence fully covered the standard. Both models (a) appropriately considered the judgements of a suitable set of educators regarding an acceptable passing standard using a similar implementation of the briefing book method, (b) utilized performance data (i.e., impact data) and candidate score profiles to inform judgements, (c) documented their process at a similarly deep and appropriate level, and (d) framed the need of each panelist to create a definition of KSAs associated with minimally qualified candidates in a similar manner. While both edTPA and CalTPA's standard-setting procedures were judged to be rigorous, both TPA models would benefit from the inclusion of validity studies that provide relationships between test performance and external criteria (e.g.,

performance evaluation data from teachers' first year of teaching) that participants could use to inform standard setting judgements during future standard-setting activities.

The procedures used by FAST were not comparable to, nor as rigorous as, those used by edTPA and CalTPA. For ADS 1(m), HumRRO rated FAST as "4" (mostly covering the Standard) because the activity related to standard-setting for the SVP and TSP was informed by teacher preparation program staff, but not the judgements of applicable non-teacher preparation staff (e.g., TK-12 teachers, TK-12 principals were not included). On *Joint Standard* 5.21, HumRRO rated FAST at a level that indicated little evidence for the Standard ("2") because the FAST model did not document the rationale for using a non-traditional standard setting process (e.g. the Body of Work, Bookmark, Briefing Book method was not used) to determine the cut-score. On *Joint Standard* 5.22 HumRRO rated FAST at a level that indicated that the evidence mostly covered the standard ("4") because the FAST model included teacher preparation program staff judgements about rubric language and performance but did not include specific candidate performances for them to consider. On *Joint Standard* 5.23, HumRRO rated FAST at a level that indicated evidence covered some of the Standard ("3") because the FAST model did not incorporate performance (i.e., impact data) as a basis for participant judgements. While an excellent review of the clarity and appropriateness of the rubrics, future FAST standard-setting activities should consider including performance data (i.e., impact data), actual candidate submissions representing a variety of performance levels, and consideration of a compensatory model in order to make it more rigorous and comparable to edTPA and CalTPA.

In conjunction with our review of TPA model scoring (Chapter 4) and the assessment designs of the TPAs, we offer one final recommendation for consideration. Given that the reliability of individual rubric scores (subscores) are not as reliable as total scores, we encourage FAST to consider adopting a compensatory scoring model, or at least limiting conditions that relate to failing candidates based on a single rubric score/subscore (e.g., failing candidates who get one '1').

Conclusion

Review of standard-setting documentation and observation of standard-setting processes indicates that CalTPA and edTPA standard-setting procedures adhere to assessment industry standards and to ADS 1(m) and are appropriate for the design and components of each assessment. The procedures of FAST, as observed during the Passing Standard Workshop, are not comparable or as rigorous as those used by edTPA and CalTPA. The procedures used by FAST constituted an appropriate approach for reviewing and revising prompts and rubric descriptors; however, the FAST model should consider including performance data (i.e., impact data), actual candidate submissions representing a variety of performance levels, and consideration of a compensatory model during its standard-setting procedures in the future.

Chapter 6: Statistical Analysis and Comparison of Score Data across TPA Models (Activity 6)

Matt Swain

Introduction

Activity 6 was conducted as an independent investigation of the veracity of Claim 6:

> The model sponsor for each TPA model conducts statistical analyses to identify differential effects in relation to candidates' race, ethnicity, language, gender or disability. Any differences are documented, and processes implemented to eliminate sources of construct-irrelevant variance.

Claim 6 stems from *Assessment Design Standard* 1(k):

> The model sponsor completes initial and periodic basic psychometric analyses to identify pedagogical assessment tasks and/or scoring rubrics that show differential effects in relation to candidates' race, ethnicity, language, gender or disability. When group pass-rate differences are found, the model sponsor investigates the potential sources of differential performance and seeks to eliminate construct-irrelevant sources of variance.

Our approach was to focus on the pass rate differences by TPA model for available demographic variables. Differential impact is a necessary, but insufficient indicator of bias. In other words, any statistical difference in pass rates may suggest, but not solely indicate, a construct-irrelevant source that explains this difference. Moreover, statistical significance may or may not indicate a practical significance of pass rates between the models. Therefore, we discuss both standardized measures of effect size and differences in the percentages of individuals passing by subgroup to provide context for statistically significant results. For this analysis, we focused on the largest credential area across all three models—the multiple subject credential. The number of candidates seeking single subject credentials was small, particularly for FAST; these numbers were too small to support subgroup analyses on demographic variables.

Method

Data Cleaning

The 2018–19 raw data for the three TPA models were provided to HumRRO. These data contained candidate rubric scores, demographic variables, and pass/fail status. In the FAST dataset, there were 103 multiple subject records. The edTPA dataset contained 1,797 multiple subject records. CalTPA had 3,809 multiple subject records for Cycle 1 and 2,927 records for Cycle 2.

As expected, there were multiple records for some candidates; therefore, additional data cleaning was necessary to handle retakes. We analyzed the data two ways: (a) the first record for each unique ID regardless of retakes and (b) the final record for each candidate after retakes. This was done in part due to the 100% pass rate for FAST after retakes. That is, because 100% of FAST candidates passed after retakes the only opportunity to observe pass

rate differences among demographic groups was by including initial attempts. Therefore, the raw data were cleaned into two datasets for each TPA model: "first attempt" and "final attempt."

For the FAST model, duplicates could not be identified. However, the data were provided in "long" format meaning there were additional variables that indicated scores for a second attempt, so retakes could be explored. There were no records designated as "Incomplete." The final analysis sample for FAST is presented in Table 6.1.

Table 6.1. FAST "First Attempt" and "Final Attempt" Sample Sizes

Assessment	First Attempt (*n* = 103)		Final Attempt (*n* = 103)	
	Number of Incomplete	Valid Records	Number of Incomplete	Records
Multiple Subject Credential	0	103	0	103

For edTPA, two duplicate records were removed. Also, several candidates completed the assessment two or three times and failed their first or second attempt. Both the first and final attempt datasets for edTPA contained 1,581 records. The final analysis sample for First Attempt and Final Attempt are presented in Table 6.2.

Table 6.2. edTPA "First Attempt" and "Final Attempt" Sample Sizes by Content Focus

Assessment Content Area Focus	First Attempt (*n* = 1,581)		Final Attempt (*n* = 1,581)	
	Number of Incomplete	Valid Records	Number of Incomplete	Valid Records
Elementary Education: Literacy with Mathematics Task 4	22	1,217	26	1,213
Elementary Education: Mathematics with Literacy Task 4	3	339	5	337
Analysis Sample	**25**	**1,556**	**31**	**1,550**

For CalTPA, we received the data such that a candidate's Cycle 1 and Cycle 2 data were provided in separate rows. Thus, it was necessary to merge by ID number to determine overall pass rates. The merged numbers for CalTPA are found in Table 6.3. Note that some of these merged numbers contained records where the candidate only completed Cycle 1 *or* Cycle 2. For CalTPA, there were several candidates identified who completed the assessment two or three times before passing on their final attempt. After removing these first and second attempt records, we retained 6,535 records in the final dataset across both cycles. There were 6,459 records in the first dataset as well, but it contained the candidate's first attempt. After concatenating the CalTPA datasets with matched valid records, there were 3,716 records in the "first attempt" dataset and 3,727 records in the "final attempt" dataset. There were several duplicate records, some of which were incomplete; these were removed. Additionally, records where only one cycle was completed (i.e., partial records) were also dropped from the analysis datasets. Partial records were more common for first attempts than final attempts. The analysis datasets for CalTPA comprised 2,743 for the first attempts and 2,808 for final attempts.

Table 6.3. CalTPA "First Attempt" and "Final Attempt" Sample Sizes by Content Area and Cycle Before and After Matching

Assessment Content Area Focus and Cycle	First Attempt			Final Attempt		
	Number of Incomplete	Valid Records	Unique IDs with Valid Score [a]	Number of Incomplete	Valid Records	Unique IDs with Valid Score [a]
Multiple Subject Cycle 1: Learning About Students and Planning Instruction (Literacy)	36	1,560	1,578	22	1,574	1,582
Multiple Subject Cycle 2: Assessment-Driven Instruction (Mathematics)	12	1,165		7	1,170	
Multiple Subject Cycle 1: Learning About Students and Planning Instruction (Mathematics)	31	2,104	2,138	7	2,128	2,145
Multiple Subject Cycle 2: Assessment-Driven Instruction (Literacy)	59	1,630		26	1,663	
Total with Duplicates		6,459	3,716		6,535	3,727
Analysis Sample with Duplicate and Partial [b] Records Removed			**2,743**			**2,808**

[a] Matched sample sizes contain some incomplete records where candidates did not complete both cycles as well as duplicate records.
[b] Partial records refer to cases where a candidate completed one cycle but not the other. This was more common for first attempts than final attempts, which is why there are more candidates in the final attempt than first attempt column.

Results

Several analyses in this chapter use a chi-square (χ^2) test of independence to determine if the observed frequencies in 2 x k tables are similar across models (where k is the number of models). Chi-square tests are affected by large sample sizes and can result in inflated Type I error rates. Therefore, due to the large sample sizes for edTPA and CalTPA, we also computed the phi coefficient (φ) as an effect size. Cramer's *V* statistic is a transformation of phi so that it is bounded by -1 and 1, like Pearson correlations, and aids in its interpretation. Both phi and *V* can be interpreted like correlation coefficients, and therefore a measure of the strength of the relation between two categorical variables (and are equivalent with 2 x 2 tables). If we had equivalent percentages of candidates in each racial category by model, we would expect a non-significant chi-square and a *V* of 0.

First, because CalTPA and edTPA have multiple "formats" (i.e., Cycle 1 focus on Math and Cycle 2 focus on Literacy or vice versa for CalTPA and two elementary education handbooks for edTPA), we investigated whether format was related to pass rates. The results of this investigation showed that format was unrelated to pass rates for CalTPA in a practical sense for

first attempts ($\chi^2(1) = 5.83$, $p = .02$, $V = -.05$) and final attempts ($\chi^2(1) = 3.80$, $p = .05$, $V = -.037$). A Cramer's V of .05 is small as would be a Pearson correlation of that size. For edTPA, pass rates were also unrelated to format in a practical sense for first attempts ($\chi^2(1) = 5.51$, $p = .02$, $V = .06$) and final attempts ($\chi^2(1) = 6.84$, $p = .009$, $V = .06$) even though the results were statistically significant. Therefore, we collapsed records across formats for both CalTPA and edTPA for all remaining analyses. Unlike edTPA and CalTPA, there is only one format of the FAST multiple subject portfolio. Thus, there was no need to check for format effects for FAST.

Comparison of Frequency Distributions

Race and/or ethnicity data were coded for all models to ensure categories were as equivalent as possible across models. Table 6.4 contains the coding scheme for race and/or ethnicity categories (hereafter referred to as "race") provided by each model. The race categories of Asian, Black, Hispanic, and White were the same across models. "Amind" in the FAST data was recoded to Native American to match the other two models; however, the n-counts were very small for this group, particularly when split by pass rates, and therefore was ultimately recoded to missing (Excluded). All other categories noted in Table 6.4 were infrequent and did not appear in all three models; therefore, they were recoded as missing. The "Final Code" column in Table 6.4 displays the race categories included in the chi-square analyses.

Comparison of Frequency Distributions

Race and/or ethnicity data were coded for all models to ensure categories were as equivalent as possible across models. Table 6.4 contains the coding scheme for race and/or ethnicity categories (hereafter referred to as "race") provided by each model. The race categories of Asian, Black, Hispanic, and White were the same across models. "Amind" in the FAST data was recoded to Native American to match the other two models; however, the n-counts were very small for this group, particularly when split by pass rates, and therefore was ultimately recoded to missing (Excluded). All other categories noted in Table 6.4 were infrequent and did not appear in all three models; therefore, they were recoded as missing. The "Final Code" column in Table 6.4 displays the race categories included in the chi-square analyses.

Table 6.4. Race Coding Scheme

FAST Categories	edTPA Categories	CalTPA Categories	Final Code
Asian	Asian	Asian	Asian
Black	Black	Black	Black
Hispanic	Hispanic	Hispanic	Hispanic
Amind	NatAmer	NatAmer	(Excluded)
N/A	Other	Other	(Excluded)
Pacif	N/A	Pac Isl	(Excluded)
White	White	White	White
Two more Unknown	Multiracial Undeclared	N/A	(Excluded)

Frequencies of the four categories in the "Final Code" column of Table 6.4 were compared across models (without considering pass status). Table 6.5 contains the frequencies of the race categories by model for first attempts. As seen in Table 6.5, the frequency distributions in the

race categories for edTPA and CalTPA are quite similar. FAST, on the other hand, differs from edTPA and CalTPA in that the majority of the candidates are identified as Hispanic as opposed to White. Thus, it was not surprising that when we statistically compared race counts by model for first attempts the result was statistically significantly different ($\chi^2(6) = 27.04$, $p < .001$). A non-significant result would mean the counts of race were similar across all models. Cramer's *V* for this analysis was .06, which is considered a small effect size. This means that while some of the racial categories' percentages were not equal between models, they were not markedly different. Given the small n-counts for FAST, the percentages are untenable and may shift with more candidates. Or, the FAST sample may be representing a different population than the other models. When excluding the FAST model, the racial category numbers are similar between CalTPA and edTPA for first attempts ($\chi^2(3) = 6.81$, $p = .07$, $V = .04$). This same analysis was conducted on final attempts and the results were similar. Between CalTPA and edTPA, race categories were similar for the final attempt; that is, the result was not statistically or practically significant ($\chi^2(3) = 6.98$, $p = .07$; $V = .04$; see Table 6.6). Dichotomizing race into "White" and "Non-White" also resulted in very similar results as keeping the race categories separate.

Table 6.5. Frequency of Race Categories by Model – First Attempt

Race	FAST		edTPA		CalTPA	
	n	%	*n*	%	*n*	%
Asian	6	6.19	149	11.19	228	9.25
Black	4	4.12	34	2.55	76	3.08
Hispanic	49	50.52	375	28.15	765	31.02
White	38	39.18	774	58.11	1,397	56.65
Total	**97** [a]	**100**	**1,332** [a]	**100**	**2,466** [a]	**100**

[a] Candidate records not in this table were recoded according to Table 6.4 as missing due to small sample size or lack of comparable race category across models.

Table 6.6. Frequency of Race Categories by Model – Final Attempt

Race	FAST [a]		edTPA		CalTPA	
	n	%	*n*	%	*n*	%
Asian	--	--	147	11.09	232	9.21
Black	--	--	33	2.49	79	3.13
Hispanic	--	--	375	28.28	783	31.07
White	--	--	771	58.14	1,426	56.59
Total	--	--	**1,326** [b]	**100**	**2,520** [b]	**100**

[a] FAST not included due to 100% pass rate for final attempt.
[b] Candidate records not in this table were recoded according to Table 6.4 as missing due to small sample size or lack of comparable race category across models.

The other demographic variable provided by all TPA models was gender. Gender of "Not Provided" was excluded from all analyses due to a small number of candidates selecting that response option. When comparing gender by model, the effect was statistically significant but practically small for first attempts ($\chi^2(2) = 8.92$, $p = .01$, $V = .05$; see Table 6.7), and not statistically significant final attempts ($\chi^2(1) = 0.066$, $p = .79$, $V = -.004$; see Table 6.8). Thus, we determined that the gender distribution across the models was very similar.

Table 6.7. Frequency of Gender Categories by Model – First Attempt

	FAST		edTPA		CalTPA	
Gender	*n*	%	*n*	%	*n*	%
Female	79	76.70	1,345	87.00	2,361	86.71
Male	24	23.30	201	13.00	362	13.29
Total	103	100	1,546 [a]	100	2,723 [a]	100

[a] Remaining candidate records were recoded from "Not Provided" to missing.

Table 6.8. Frequency of Gender Categories by Model – Final Attempt

	FAST [a]		edTPA		CalTPA	
Gender	*n*	%	*n*	%	*n*	%
Female	--	--	1,339	87.00	2,418	86.73
Male	--	--	200	13.00	370	13.27
Total	--	--	1,539 [b]	100	2,788 [b]	100

[a] FAST not included due to 100% pass rate for final attempt.
[b] Candidate records not in this table were recoded according to Table 6.4 as missing due to small sample size or lack of comparable race category across models.

Comparison of Pass Rates

Passing status was provided in the data files by the model sponsors and verified by HumRRO. For FAST, in order to pass a candidate had to score a "2" or higher on each rubric, which equates to a cut score of 20 across the 10 rubrics. For edTPA, a score of 49 (out of 90) was set as the cut score (there are 18 rubrics for edTPA multiple subject). CalTPA designated a candidate as "passing" with a final cut score of 19 (out of 40) on Cycle 1 with one score of 1 allowed (out of 8 rubrics) and a final cut score of 21 (out of 45) on Cycle 2 with one score of 1 allowed (out of 9 rubrics).

Pass Rates Overall

First, we examined the pass rates by model overall. For the first attempts, the pass rates are statistically and practically different by model ($\chi^2(2) = 57.98$, $p < .0001$, $V = .11$; see Table 6.9). Considering the frequencies within each model, for first attempts a candidate is slightly more likely to pass CalTPA than edTPA and much more likely to pass CalTPA than FAST. For final attempts, 100% of FAST candidates passed. Because there is no variance in FAST pass rates on final attempts, the FAST model is excluded from further analyses of pass rate differences on final attempts. The difference between CalTPA and edTPA still exists when considering candidates' final attempts and to a slightly larger extent, ($\chi^2(1) = 88.17$, $p < .0001$, $V = -.14$; see Table 6.10).

Table 6.9. Pass Rates Overall by TPA Model – First Attempt

FAST			edTPA			CalTPA		
Fail	Pass	Total	Fail	Pass	Total	Fail	Pass	Total
14	89	103	86	1,470	1,556	63	2,680	2,743
13.59%	86.41%	100%	5.53%	94.47%	100%	2.30%	97.70%	100%

Table 6.10. Pass Rates Overall by TPA Model – Final Attempt

FAST			edTPA			CalTPA		
Fail	Pass	Total	Fail	Pass	Total	Fail	Pass	Total
0	103	103	80	1,470	1,550	20	2,788	2,808
00.00%	100.00%	100.00%	5.16%	94.84%	100%	0.71%	99.29%	100%

Pass Rates by Race

In the prior analyses, chi-square tests of independence have been sufficient to compare 3 x k tables, where k is the number of levels for one categorical variable (e.g., model x gender). When we add passing status, we then have 2 x 3 x k tables (i.e., passing status by model by demographic variable). For these comparisons, we used an extension of the test of independence called the Cochran-Mantel-Haenszel (CMH) test. This statistic is also chi-square distributed but does not have an effect size (like Cramer's *V*) so we rely on the differences in pass rates for any statistically significant results and look at the chi-square test of independence when controlling for model (i.e., within model differences).

Pass rates across models appeared to differ by race for first attempt ($\chi^2(1) = 6.60$, $p = .01$) and final attempt ($\chi^2(1) = 5.75$, $p = .02$). However, pass rates by race within model were not statistically or practically significant for first attempts ($\chi^2(3) = 6.57$, $p = 0.09$, $V = .07$ and $\chi^2(3) = 5.83$, $p = .12$, $V = .05$) for edTPA and CalTPA, respectively. The same result was true for final attempts ($\chi^2(3) = 4.68$, $p = .20$, $V = .06$ and $\chi^2(3) = 6.57$, $p = .09$, $V = .05$) for edTPA and CalTPA, respectively. Moreover, some racial categories had small sample sizes (e.g., Black), which makes the pass rates for those groups statistically tenuous and difficult to compare across models (Renter, Higgins, & Sargeant, 2000). Regardless, the pass rates in Tables 6.11 and 6.12 show the similarity in pass rates within model and across racial groups.

Table 6.11. Pass Rates by Race and TPA Model – First Attempt

Race	edTPA			CalTPA		
	Fail	Pass	Total	Fail	Pass	Total
Asian	12	137	149	5	223	228
	8.05%	91.95%	100%	2.19%	97.81%	100%
Black	4	30	34	2	74	76
	11.76%	88.24%	100%	2.63%	97.37%	100%
Hispanic	23	352	375	24	741	765
	6.13%	93.87%	100%	3.14%	96.86%	100%
White	34	740	774	22	1,375	1,397
	4.39%	95.61%	100%	1.57%	98.43%	100%
Total	73	1,259	1,332	53	2,413	2,466

Note. FAST was not included due to a lack of similarity in race distribution with the other models. See Table 6.5.

Table 6.12. Pass Rates by Race and TPA Model – Final Attempt

Race	edTPA			CalTPA		
	Fail	Pass	Total	Fail	Pass	Total
Asian	10	137	147	2	230	232
	6.80%	93.20%	100%	0.86%	99.14%	100%
Black	3	30	33	1	78	79
	9.09%	90.91%	100%	1.27%	98.73%	100%
Hispanic	23	352	375	10	773	783
	6.13%	93.87%	100%	1.28%	98.72%	100%
White	31	740	771	5	1,421	1,426
	4.02%	95.98%	100%	0.35%	99.65%	100%
Total	67	1,259	1,326	18	2,502	2520

Note. FAST was not included due to a lack of similarity in race distribution with the other models (Table 6.5.) *and* due to the 100% pass rate for final attempts.

Pass Rates by Gender

Controlling for model, passing rates differed by gender ($\chi^2(1) = 13.05$, $p < .001$) for first attempts (see Table 6.13). Females had slightly higher pass rates than males on all models. Looking within model, the pass rates between the genders for CalTPA was statistically significant but practically small ($\chi^2(1) = 13.06$, $p < .001$; $V = -.07$). This difference was not statistically or practically significant for edTPA ($\chi^2(1) = 2.87$, $p = .09$; $V = -.04$) or FAST ($\chi^2(1) = 0.25$, $p = .62$; $V = -.05$). The difference for CalTPA was likely due to the larger sample size for CalTPA because the difference in pass rates between the genders was similar across models (≈4% for FAST and ≈3% for edTPA). However, these differences are practically small (see Table 6.13).

Table 6.13. Pass Rates by Gender and TPA Model – First Attempt

Gender	FAST			edTPA			CalTPA		
	Fail	Pass	Total	Fail	Pass	Total	Fail	Pass	Total
Female	10	69	79	68	1,277	1,345	45	2,316	2,361
	12.66%	87.34%	100%	5.06%	94.94%	100%	1.91%	98.09%	100%
Male	4	20	24	16	185	201	18	344	362
	16.67%	83.33%	100%	7.96%	92.04%	100%	4.97%	95.03%	100%
Total	14	89	103	20	1,526	1,546	63	2,660	2,723

For the final attempt analysis, pass rates differed by gender after controlling for model ($\chi^2(1) = 6.58$, $p = .01$). This means that one model has a difference between the genders (see Table 6.14). Looking within model, the pass rate difference between the males and females for CalTPA was statistically significant but practically small ($\chi^2(1) = 4.90$, $p = .03$; $V = -.04$). However, this difference was not statistically or practically significant for edTPA ($\chi^2(1) = 3.01$,

p = .08; *V* = -.04). The difference for CalTPA was practically very small considering Cramer's *V* as well as the observed difference of ≈1% and is likely driven by large sample sizes (see Table 6.14).

Table 6.14. Pass Rates by Gender and TPA Model – Final Attempt

Gender	FAST [a]			edTPA			CalTPA		
	Fail	Pass	Total	Fail	Pass	Total	Fail	Pass	Total
Female	--	--	--	62	1,277	1339	14	2,404	2,418
	--	--	--	4.63%	95.37%	100%	0.58%	99.42%	100%
Male	--	--	--	15	185	200	6	364	370
	--	--	--	7.50%	92.50%	100%	1.62%	98.38%	100%
Total	--	--	--	77	1,462	1,539	20	2,768	2,788

[a] FAST not included due to 100% pass rate for final attempt.

Comparison of Mean Differences

As a final analysis, we compared mean total scores (i.e., sums of rubric scores) by race and gender. As noted above, some of the results yielded statistically significant differences, but the actual percentage differences appeared to be practically small. When sample sizes are large, as tended to be the case with edTPA and CalTPA, very modest differences may meet the criterion of statistical significance, but not be practically significant. A measure of effect size, such as Cohen's *d*, can help to inform whether such differences are practically significant. Thus, we investigated the magnitude of differences in mean total scores (i.e., effect size) using Cohen's *d.* Cohen's *d* was computed using the following formula:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}$$

Where $\bar{x}_1$ was the mean of group 1; $\bar{x}_2$ was the mean of group 2; s_1^2 was the variance of group 1; and s_2^2 was the variance of group 2.

Mean Scores by Race

We compared mean scores between White and Non-White candidates. Comparison of racial groups for the first attempts are presented in Table 6.15 and for the final attempts in Table 6.16. White candidates scored higher than their Non-White counterparts in almost all comparisons across models, although the magnitude of the differences were near zero or small, using Cohen's (1988) guidelines of ≤ .20 as small, .50 as moderate, and .80 as large. We did not compute differences between Asian and White and between Black and White for FAST due to the small sample sizes (*n* < 10). There was a relatively large effect size difference between Hispanic and White candidates on FAST; Hispanic candidates earned almost three fewer points, on average, than White candidates on first attempts and just over two points fewer, on average, on final attempts.

Table 6.15. Mean Total Scores by Race – First Attempt

Model	Race	Non-White			White			
		n	Mean	*SD*	*n*	Mean	*SD*	Cohen's *d*
FAST [a]	Asian	--	--	--	38	27.82	3.83	--
	Black	--	--	--	38	27.82	3.83	--
	Hispanic	49	24.96	3.25	38	27.82	3.83	-0.81
edTPA	Asian	149	55.59	5.54	774	55.55	4.99	0.01
	Black	34	54.56	4.51	774	55.55	4.99	-0.20
	Hispanic	375	54.59	5.47	774	55.55	4.99	-0.19
CalTPA	Asian	228	49.47	5.99	1,397	49.44	6.49	<0.01
	Black	76	48.20	5.99	1,397	49.44	6.49	-0.19
	Hispanic	765	48.94	6.39	1,397	49.44	6.49	-0.08

[a] Results suppressed for groups with $n < 10$.

Table 6.16. Mean Total Scores by Race – Final Attempt

Model	Race	Non-White			White			
		n	Mean	*SD*	*n*	Mean	*SD*	Cohen's *d*
FAST [a]	Asian	--	--	--	35	28.17	3.75	--
	Black	--	--	--	35	28.17	3.75	--
	Hispanic	40	25.78	3.00	35	28.17	3.75	-0.71
edTPA	Asian	147	55.61	5.74	771	55.52	5.15	0.02
	Black	33	54.67	4.67	771	55.52	5.15	-0.17
	Hispanic	375	54.55	5.57	771	55.52	5.15	-0.18
CalTPA	Asian	232	49.53	5.93	1,426	49.53	6.37	<0.01
	Black	79	48.20	5.84	1,426	49.53	6.37	-0.21
	Hispanic	783	49.06	6.25	1,426	49.53	6.37	-0.07

[a] Results suppressed for groups with $n < 10$.

Mean Scores by Gender

Comparison of mean scores by gender for first attempts are presented in Table 6.17 and for the final attempts in Table 6.18. The effect sizes were small using Cohen's (1988) guidelines of $\leq .20$ as small, .50 as moderate, and .80 as large. However, effect sizes should be interpreted within context to determine meaning beyond these guidelines. The mean difference between gender groups was less than one point and no more than two points, which appears small. Male candidates scored higher than female candidates on FAST, but this difference is flipped for CalTPA and edTPA; however, the effect sizes are all small.

Table 6.17. Mean Total Scores by Gender – First Attempt

Model	Female			Male			
	n	Mean	*SD*	*n*	Mean	*SD*	Cohen's *d*
FAST	79	25.89	3.47	24	26.46	4.72	-0.15
edTPA	1,345	55.34	5.21	201	54.31	5.24	0.20
CalTPA	2,361	49.46	6.39	362	47.87	6.56	0.25

Table 6.18. Mean Total Scores by Gender – Final Attempt

Model	Female			Male			
	n	Mean	*SD*	*n*	Mean	*SD*	Cohen's *d*
FAST	69	26.49	3.22	20	27.70	3.89	-0.36
edTPA	1,339	55.31	5.37	200	54.32	5.30	0.19
CalTPA	2,418	49.56	6.28	370	48.15	6.27	0.22

Discussion and Conclusion

We compared pass rates and total scores across models by race and gender. These analyses were conducted as an independent investigation of the veracity of Claim 6, which directs model sponsors to identify "…*differential effects in relation to candidates' race, ethnicity, language, gender or disability*." We focused on race and gender as these data were available from all three models. However, more indicators, notably language and disability, should be collected by model sponsors to fully explore and address Claim 6. We also focused on the multiple subject credential given the small number of candidates submitting single subject portfolios, particularly for FAST. Because the multiple subject credential is the most frequently sought credential it afforded the greatest opportunity to investigate demographic differences among subgroups. However, the pattern of subgroup differences in pass rates may be very different depending on the credential area being investigated (e.g., single subject mathematics). Thus, the findings in this activity should not be extrapolated to other credential areas.

We found that candidates had similar pass rates by gender across the three models. Although there were some statistically significant results, the effect sizes were small or very small and when examining actual percentage differences, the practical difference was non-substantial. Moreover, mean total scores were similar by gender for all three models resulting in small standardized effect sizes; however, it is worth noting that the direction of these small differences were such that females tended to score slightly higher on edTPA and CalTPA, whereas males tended to score slightly higher on FAST, although the samples sizes were particularly small for FAST and this pattern may shift with more candidates or different FAST cohorts.

Findings comparing models by race were more complex. Before looking at pass rates, the number of candidates in each racial category were compared across the three models and resulted in our largest effect size in this chapter (V = -.14), although still small. This result indicated that the racial categories of the FAST sample were disproportionate to the other two models. Given that the racial composition of the FAST sample differed from the racial composition of the edTPA and CalTPA samples, FAST was removed from the comparison of pass rates by model analysis. After removing FAST, pass rates did not differ meaningfully between CalTPA and edTPA by race categories.

In conclusion, we found no evidence to suggest substantive differences in pass rates for males and females within TPA models. Moreover, the pattern of pass rates for males and females was comparable across models. In addition, when we examined differences in mean total scores the magnitude of the differences between males and females were similarly small for all three models. These findings support the claim that there are no differential effects in relation to candidates' gender (Claim 6). The findings also indicate that the pass rates for the various race categories were similar both within and across models for edTPA and CalTPA, thereby lending support to the claim that there are no differential effects in relation to candidates' race (Claim 6). Comparisons of mean total scores showed no notable differences among racial groups for any of the models except that White candidates tended to have higher mean total scores on FAST than Hispanics candidates, but all candidates ultimately passed the FAST TPA and, thus, the differences in mean total scores did not translate to differences in pass rates. It's worth noting that while final pass rates for edTPA and CalTPA where not 100% like they were for FAST, they were, nonetheless, very high, particularly for CalTPA (around 99%). Again, these results are based on the multiple-subject credential only and should be revisited as more data (for multiple subject and for other credential areas) becomes available. Also, the models should collect data on other demographic variables—notably, language and disability—so that ADS 1(k) can be fully investigated.

Chapter 7: Comparison of TPA Models to a Common Criterion (Activity 7)

Andrea Sinclair, Sunny Becker, Justin Paulsen, Wade Buckland, & Arthur Thacker

Introduction

The final activity represents an innovative and informative method for investigating the ultimate question of comparability across TPA models. No group of candidates completes more than one TPA nor are there any common "items" across TPA models. Consequently, there is no referent test or common criterion that can be accessed to compare TPA models. Thus, in the previously described activities, the question of comparability was addressed by examining the internal functioning of each TPA model and qualitative indicators of comparability based on content, strength of evidence, stakeholder perceptions, etc., all of which are important, but none of which would meet many definitions of true comparability. Consequently, for Activity 7 we developed a common external referent on which all the TPAs can be directly compared. While this does not represent what many researchers would consider a true comparability study, whereby the tests would be considered interchangeable for generating candidates scores and passing status, it does give us a common referent on which we can examine candidate performance across the TPAs.

We used the results from the previous activities, particularly Activity 2—the content validity investigation, to identify a list of TPE elements that are assessed in substantively the same way across the TPA models (e.g., all models require candidates to submit video and commentary as evidence for TPE element 1.8) and for which all the models measure the full depth and breadth of those TPE elements (see Tables 2.1 - 2.6 in Chapter 2 for the TPE elements mapped to each model). We then developed a "Common Rubric" to measure those TPE elements. Trained and calibrated assessors scored a representative sample of candidate submissions from each model using the Common Rubric. We then conducted comparability analyses across TPA models using the scores on the Common Rubric as a referent. That is, we correlated scores on the Common Rubric with operational scores on each model's TPA-specific rubric. We also compared cut scores and the percentage of pass/fail indications on the TPA-specific rubrics with the percentage of interpolated pass/fail indications on the Common Rubric. The benefit of this activity is that the same candidate submissions were rated on a common criterion. The details of the methodology are presented next.

Method

Our methodology consisted of four tasks: a) Select Representative Sample of Scored Candidate Submissions, b) Identify Common Scoring Staff, c) Develop Common Rubric, and d) Score Candidate Submissions Using Common Rubric.

Select Representative Sample of Scored Candidate Submissions

The evaluation design required that 30 multiple subject credential candidate submissions from each model be scored using the Common Rubric and that these submissions represent a diverse level of quality. It was beyond the scope of this study to conduct this activity for all credential areas. Thus, in discussion with the technical advisory committee (TAC), we decided to conduct this activity for the most commonly sought credential—the multiple subject credential. Official scores (as assigned operationally by each TPA model) were used as a proxy for level of quality. In total, 40 submissions were collected from each model to accommodate common scoring of 30 submissions per model, as well as rubric tryouts, calibration, and substitutions if needed.

The HumRRO Project Director requested raw score data for all 2018–19 candidate portfolios from model sponsors. Based on the total first attempt raw scores, she identified a stratified random sample of 40 candidate portfolio IDs per model. First attempt scores were used so that there would be more variance (i.e., more to predict) in the data.[34] The stratified random sample was selected as follows. Data were sorted by total score and stratified into quintiles (i.e., fifths). Eight candidate IDs were randomly selected from each stratum, for a total of 40 candidate portfolios per model. The edTPA portfolios are structured as Literacy with Math Task 4 or Math with Literacy Task 4; 20 candidate portfolios were selected for each. Similarly, CalTPA submissions comprise two pieces: Cycle 1 and Cycle 2. An individual candidate may submit one of two formats: Cycle 1 with Math focus and Cycle 2 with Literature focus, or vice versa. CalTPA scores for Cycles 1 and 2 were combined to obtain a total score for quintile assignment, and 20 candidate portfolios were identified for each format.

The HumRRO Project Director sent the edTPA and CalTPA model sponsors the list of 40 candidate IDs and asked them to upload the associated portfolios to HumRRO's ftp site. The FAST model sponsor, however, selected the stratified random sample of 40 portfolios IDs, and submitted those materials to the ftp site.

A HumRRO staff member, not involved with scoring the submissions, randomized the order of the submissions within each model, based on total operational scores. As a secondary step to ensure balance, the order of edTPA portfolios were randomized by Math versus Literacy focus and CalTPA submissions were randomized in terms of Literacy versus Mathematics focus of the cycles. Because HumRRO scoring was blind, operational scores were stripped from the materials provided to HumRRO scorers.

Identify Common Scoring Staff

The common scoring task required two scorers with specific knowledge of the design of the three TPA models, as well as substantial expertise in the development and evaluation of rubrics, scoring processes, and human scorer training procedures. One HumRRO researcher and one emeritus HumRRO researcher served in this capacity (hereafter, *HumRRO scorers*). These individuals attended the training events identified in Table 7.1. In addition, these individuals reviewed and evaluated model materials with respect to scoring as described in Chapter 4 of this report, *Comparison of Scoring Rubrics, Score Reports, and Rater Training*.

Table 7.1. Scorer Training and Calibration Observations Attended by HumRRO Scorers

Scorer Training Event	Date(s)	Mode / Location
FAST		
Mathematics TSP Assessor Training and Calibration	4/9/18	Onsite/Fresno, CA
Multiple Subject TSP Assessor Training & Calibration	4/9/18	Onsite/Fresno, CA
edTPA		
Online Training	May 2018	Remote/Asynchronous
Calibration	None	None
CalTPA		
Assessor Orientation	1/19/18	Webinar
Assessor Orientation	2/13/18	Webinar
Cycle 1 Assessor Training & Calibration (southern)	3/20/18	Onsite/San Bernardino, CA
Cycle 2 Assessor Training & Calibration (northern)	4/11/18	Onsite/Sacramento, CA
Cycle 1 & 2 Assessor Training & Calibration (southern)	11/26-27/18	Pomona, CA

[34] We know from the analysis in Chapter 6 that first attempt and final attempt pass rates were similarly high. Thus, we would not expect substantively different results from those reported here had we requested final attempt portfolios.

The HumRRO scorers had substantial scoring expertise preceding this study. One scorer has 12 years of experience evaluating scoring processes, procedures, and outcomes of large-scale student achievement and teacher certification measures. This work includes evaluations of the National Assessment of Educational Progress, California Assessment of Student Performance and Progress (CAASPP), and New York teacher certification exams. Studies included evaluation of scorer training, scorer monitoring, documentation, and use of second scores and validity papers. His research focus has been on the development of materials and adherence to proven procedures and standards that promote reliable ratings. In addition, he has provided recommendations to Pearson, ETS, and Measured Progress on how to improve their distributed scoring platforms.

The other scorer led and participated in scoring studies for over 15 years, including development of rubrics and scoring processes and evaluations of rubrics and processes developed by others. She led a study of hand scoring processes by two vendors for the California Assessment of Student Performance and Progress (CAASPP) as part of the independent evaluation contract on behalf of the California Department of Education (CDE). She previously led a review of the scoring operations for the New York State Teacher Certification Examinations (NYSTCE) program, and directed HumRRO's quality assurance work for the National Assessment of Educational Progress (NAEP). The NAEP activities included reviewing scoring materials and designing and implementing an extensive quality assurance process for NAEP scoring, and training several researchers to conduct this process annually. This quality assurance process monitored and provided actionable feedback on NAEP scoring processes; for example, a recommendation to embed expert-scored papers into the scoring process to improve scoring consistency was implemented operationally. She participated in a study to examine the comparability of Partnership for Assessment of Readiness for College and Careers (PARCC) scores across states, assessment forms, and scoring methods. She also served as a consultant for the development of performance-based assessments on a U.S. Department of Education Catalyst Grant, Preparing Tomorrow's Teachers to Use Technology (PT^3) for the Maryland State Department of Education. This grant was intended to ensure that teacher candidates are prepared to use technology in the classroom for teaching and learning. Duties as the team leader included working with a consortium of university-level educators to design assessments, leading the development of scoring rubrics, coordinating consistent application of scoring, organizing shadow-scoring activities, preparing implementation materials for distribution throughout the state, and evaluating field test results.

The HumRRO Project Director and the HumRRO Senior Technical Advisor overseeing this comparability study both have experience scoring teacher portfolios, specifically teacher portfolios for the National Board for Professional Teaching Standards (NBPTS). They provided oversight to the HumRRO scorers scoring the TPA portfolios.

Develop Common Rubric

A Common Rubric was central to the evaluation design, so that candidate submissions from each of the three models could be scored based on the same expectations. That is, an external criterion measure was designed to assess the TPE elements common to (i.e., measured by) all three models. The steps for developing the Common Rubric are outlined below.

1. In April 2018, a four-day workshop was conducted with two, 7-person panels of teacher preparation experts. The two panels independently mapped TPE elements to each component of each TPA. The experts' evaluations extended beyond a simple evaluation

of coverage of TPE elements by TPAs. They also identified the type of evidence required by each model to assess each TPE element and an evaluation of how thoroughly each model assesses the KSAs specified by each TPE element. At the end of the workshop, the two panels produced a cross-validated matrix (one for each TPA model) mapping each TPE element to each component of the TPA according to the type of evidence required by the TPA component (e.g., lesson plan, video, reflection), along with a rating indicating how thoroughly the collective set of evidence required by the TPA assesses the KSAs in each TPE element. This matrix served as the blueprint for developing the Common Rubric. Full details on the content validity workshop can be found in Chapter 2 of the Year 1 report (Sinclair & Thacker, 2018).

2. HumRRO researchers hosted a webinar with five teacher preparation experts on August 14-16, 2019 as a follow-up to the April 2018 workshop. The five experts who participated in the August 2019 webinar included experts on each of the three models. All five experts had participated in the April 2018 workshop. During the August 2019 webinar, CalTPA and edTPA model sponsors presented changes made to their models since spring 2018 (if any). Because the FAST representative was unavailable, the HumRRO Project Director described changes to the FAST model. The experts considered this information as they reviewed and confirmed/revised the spring 2018 TPE-to-TPA mappings. In the webinar, participants made limited, targeted updates to the original mappings. The full details on this webinar can be found in Chapter 2 of this report.

3. The HumRRO Project Director used the (a) TPE-to-TPA mappings from the aforementioned steps and (b) individual model scoring materials (namely, the Performance Assessment Guides/Handbooks/Manual) to draft a Common Rubric. Three mandatory criteria qualified a TPE element to be included on the Common Rubric: (a) the TPE had to be linked to one or more components of each TPA; (b) evidence requirements overlapped across models for that TPE element (e.g., all three models require video and commentary evidence for that TPE element) and (c) all three models measured the full depth and breadth of the TPE element. There were six TPE elements that met all of these criteria for the three models. Three additional TPE elements met the criteria for the CalTPA and FAST models but not the edTPA model.[35] Table 7.2 lists the six TPE elements assessed by the Common Rubric.

[35] The three additional TPE elements that met these criteria for CalTPA and FAST were TPEs 2.5, 4.1, and 5.2.

Table 7.2. TPE Elements Included on Common Rubric

TPE	TPE Element
TPE 1: Engaging and Supporting All Students in Learning	1.8. Monitor student learning and adjust instruction while teaching so that students continue to be actively engaged in learning.
TPE 3: Understanding and Organizing Subject Matter for Student Learning	3.1. Demonstrate knowledge of subject matter, including the adopted California State Standards and curriculum frameworks.
TPE 3: Understanding and Organizing Subject Matter for Student Learning	3.2. Use knowledge about students and learning goals to organize the curriculum to facilitate student understanding of subject matter and make accommodations and/or modifications as needed to promote student access to the curriculum.
TPE 3: Understanding and Organizing Subject Matter for Student Learning	3.5. Adapt subject matter curriculum, organization, and planning to support the acquisition and use of academic language within learning activities to promote the subject matter knowledge of all students, including the full range of English learners, Standard English learners, students with disabilities, and students with other learning needs in the least restrictive environment.
TPE 5: Assessing Student Learning	5.1. Apply knowledge of the purposes, characteristics, and appropriate uses of different types of assessments (e.g., diagnostic, informal, formal, progress-monitoring, formative, summative, and performance) to design and administer classroom assessments, including use of scoring rubrics.
TPE 6: Developing as a Professional Educator	6.1. Reflect on their own teaching practice and level of subject matter and pedagogical knowledge to plan and implement instruction that can improve student learning.

1. Two HumRRO scorers, whose qualifications were described previously, conducted rubric tryouts. The HumRRO scorers reviewed one candidate submission from each model individually, for a total of three portfolios, and then met to discuss and assign scores on each rubric. They maintained notes regarding any problematic or unclear rubric elements and suggested revisions to ensure the rubric contents were clear, consistent, and represented a logical progression from score levels one through five.

2. The HumRRO Project Director made final revisions to the draft Common Rubric. The final Common Rubric is in Appendix 7.A.[36]

Score Candidate Submissions Using Common Rubric

HumRRO scorers next began official common scoring, including calibration, independent scoring, monitoring consistency, and reaching consensus when individual scores differed. In concert with industry-acknowledged best practices, the HumRRO scoring process was

[36] Appendices for this report are in Volume II: Appendices.

compatible with the *Assessment Design Standards* (ADS) related to scorer training and qualifications. Specifically, elements of two ADS were relevant: Assessment Designed for Validity and Fairness (ASD 1) and Assessment Designed for Reliability and Fairness (ASD 2). Table 7.3 lists the relevant excerpts from these Standards and commentary on how each was addressed in common scoring.

Table 7.3. Assessment Design Standard Elements in the Context of Common Scoring

Assessment Design Standard Excerpt	Implications for Common Scoring
ADS 1(h) The model sponsor develops 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 or any other bias that are not likely to affect job effectiveness and/or student learning.	HumRRO scorers participated in model sponsor training that included admonitions to ignore candidate factors unrelated to pedagogical competence, job effectiveness, and student learning. This approach is consistent with HumRRO scorers' previous experience in other scoring programs. HumRRO's Common Rubric minimized the candidate factors that are not clearly related to pedagogical competence.
ADS 2(c) The assessor training program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and the multi-level scoring rubrics.	HumRRO scorers participated in model sponsor training that included familiarity with the TPEs and pedagogical assessment tasks and reviewed these materials at the outset of common scoring activities. HumRRO scorers were very familiar with the Common Rubric.
ADS 2(c) The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy and calibration in relation to the scoring rubrics associated with the task.	In the common scoring environment, HumRRO scorers self-monitored the results of task-based scoring trials to ensure calibration. Ongoing reconciliation of double scoring on a twice weekly basis ensured sustained consistency and accuracy.
ADS 2(c) The model sponsor uses only assessors who successfully calibrate during the required TPA model assessor training sequence.	HumRRO scorers successfully calibrated prior to embarking on official common scoring.
ADS 2(e) The model sponsor provides a detailed plan for establishing and maintaining scorer accuracy and inter-rater reliability during field testing and operational administration of the assessment.	Scoring accuracy and consistency were evaluated twice weekly, upon completion of double scoring of six candidate submissions (two from each model).
ADS 2(e) The scoring process conducted by the model sponsor to assure the reliability and validity of candidate outcomes on the assessment may include, for example, regular auditing, selective back reading, and double scoring of candidate responses near the cut score by the qualified, calibrated scorers trained by the model sponsor.	Every candidate submission was double scored and scores were reconciled routinely. A method was in place to address any discrepancies that could not be resolved by the two HumRRO scorers. Because of the limited scope of the study, other methods of reliability and validity assurance (e.g., back reading, and validity papers) were not available.

The HumRRO scoring proceeded with calibration, independent double scoring, and reconciliation.

1. HumRRO scorers conducted a calibration exercise using one candidate submission from each model, for a total of three calibration portfolios. Each scorer independently reviewed one submission and assigned a score level on each rubric. The pair met to discuss each rubric with respect to the candidate and agree upon a consensus score before moving on to the next candidate's submission. All original individual ratings were exact or adjacent, indicating a sufficiently calibrated understanding of the rubric.

2. HumRRO scorers then began official common scoring. All candidate submissions were double scored and reconciled to maximize rigor. In order to protect against scoring drift and to minimize the effects of memory loss regarding candidate submissions over time, the HumRRO scorers met frequently to compare scores and resolve differences. They scored 12 submissions per week (four from each model) and reconciled any non-adjacent scores twice weekly. Specifically, once both HumRRO scorers completed scoring 6 candidate submissions, one scorer merged the two Excel files to include both sets of independent ratings. She then produced a third Excel row for each candidate with a consensus score, following these rules:

    a. If both HumRRO scorers assigned the same score, this score was used as the official score of record, flagged as an Exact Match.

    b. If the HumRRO scorers assigned different, but adjacent scores (e.g., 2 and 3 ,3 and 4, 4 and 5), the higher score was used as the official score of record and it was flagged as an Adjacent Score.

    c. If the HumRRO scorers assigned different scores that were at least 2 points apart, the HumRRO scorers met to discuss their ratings and decide upon an official score of record. This official score of record was flagged as a Nonadjacent Score. A process was in place to include the Project Director to settle differences that HumRRO scorers could not resolve but this option was never invoked.

    Table 7.4 summarizes the agreement rates for each model and overall. Each row represents one meeting to resolve scores for six candidate submissions. Percentages in this table include the percentage of scores for which the two HumRRO scorers assigned identical scores ("exact" agreement), the cumulative rates in which scores were identical or adjacent ("exact/adjacent"), and the percentage of scores at least 2 points apart ("nonadjacent"). Overall, 96 percent of scores were exact or adjacent, indicating acceptable agreement between scorers. Rates for each model were similar, and no individual model fell below 94 percent exact/adjacent.

    After completion of scoring, to the extent possible, we compared common scoring agreement rates to the agreement rates for operational scoring for each model. However, they were not directly comparable. All candidate submissions were double scored in the common scoring process, to maximize reliability and consistency. This level of checking is not always feasible for operational programs; for example, edTPA and CalTPA conduct double scoring only on portfolios with scores close to the passing standard, as extra confirmation that the pass/no pass decision is valid. We describe here the comparisons made between common scoring agreement rates and each model's results.

- While 100 percent of common scored portfolios were double scored, edTPA double scoring is focused on portfolios that are close to the passing standard, approximately 10 percent of the sample. The edTPA documentation does not specify a standard for agreement rates on double scoring but does indicate that standards for back reading conferences are similar to qualification standards. The edTPA qualification standards for "Typical Fields(National) – 15 rubrics" require scorers to achieve 46.7 percent exact matches to the predetermined scores on qualification portfolios and 93.3 percent exact plus adjacent scores. The common scoring agreement rates on edTPA portfolios were 57 percent and 96 percent respectively, somewhat higher than the edTPA rates.

- CalTPA also limited double scoring to portfolios with scores close to the passing standard. Double scoring agreement standards are not specified in the CalTPA documentation, but it does say Validity scoring uses the same standards as Calibration scoring. CalTPA qualification standards are cycle dependent. Cycle 1 scores must be at least 37.5 percent exact and 87.5 percent adjacent and Cycle 2 scores must be at least 44.4 percent exact and 88.9 percent adjacent. The common scoring agreement rates exceeded both of these standards, at 59 percent and 97 percent, respectively.

- FAST does not use periodic checks of the statistical properties of scores assigned by individual scorers during a scoring session to provide feedback to the scorers during the scoring window. Due to infrastructure constraints, such a process is difficult. Fifteen percent of portfolios are double scored at the end of the administration window to document interrater reliability. FAST documentation offers no agreement standard other than calibration. Scoring procedures evolved and improved over time. In 2018-19 after group training on two exemplars, first-time assessors were required to independently score a third exemplar on which they had to meet a performance threshold; to meet the calibration threshold scorers needed exact matches on at least four of the seven TSP rubrics. Scores that were not exact matches had to be at least one score point adjacent to the correct score point. In other words, the most relevant FAST standard was 57 percent exact and 100 percent exact/adjacent. Common scoring yielded rates of 59 percent and 94 percent, respectively. Common scoring fell short on the rates of FAST exact/adjacent scores, only. However, one complicating factor is that FAST operational scores were on a 4-level rubric, while the Common Rubric used 5 levels. The more fine-grained, larger number of levels on the Common Rubric affords greater opportunity for non-exact scores.

Table 7.4. HumRRO Scorer Agreement Rates by Model and Overall

Consensus Date	RATES											
	edTPA			CalTPA			FAST			TOTAL		
	Exact%	Exact/ Adjacent%	Non-Adjacent%	Exact%	Exact/ Adjacent%	Non-Adjacent%	Exact%	Exact/ Adjacent%	Non-Adjacent%	Exact%	Exact/ Adjacent%	Non-Adjacent%
9/13/2019	33.0	83.0	17.0	61.0	94.0	6.0	28.0	61.0	39.0	42.0	79.0	21%
9/18/2019	42.0	83.0	17.0	33.0	94.0	6.0	28.0	100.0	0.0	33.0	94.0	6%
9/20/2019	58.0	100.0	0.0	44.0	78.0	22.0	67.0	100.0	0.0	56.0	92.0	8%
9/25/2019	58.0	100.0	0.0	78.0	100.0	0.0	44.0	89.0	11.0	60.0	96.0	4%
9/27/2019	58.0	100.0	0.0	61.0	89.0	11.0	61.0	100.0	0.0	60.0	96.0	4%
10/3/2019	50.0	100.0	0.0	67.0	100.0	0.0	61.0	100.0	0.0	60.0	100.0	0%
10/4/2019	75.0	92.0	8.0	56.0	100.0	0.0	33.0	72.0	28.0	52.0	88.0	13%
10/9/2019	42.0	100.0	0.0	61.0	100.0	0.0	56.0	94.0	6.0	54.0	98.0	2%
10/11/2019	67.0	100.0	0.0	78.0	100.0	0.0	67.0	100.0	0.0	71.0	100.0	0.0
10/16/2019	58.0	100.0	0.0	50.0	100.0	0.0	78.0	100.0	0.0	63.0	100.0	0.0
10/18/2019	42.0	100.0	0.0	56.0	100.0	0.0	67.0	100.0	0.0	56.0	100.0	0.0
10/23/2019	42.0	83.0	17.0	67.0	100.0	0.0	72.0	94.0	6.0	63.0	94.0	6.0
10/25/2019	67.0	100.0	0.0	67.0	100.0	0.0	72.0	100.0	0.0	69.0	100.0	0.0
10/29/2019	83.0	100.0	0.0	44.0	100.0	0.0	72.0	100.0	0.0	65.0	100.0	0.0
10/30/2019	83.0	100.0	0.0	67.0	94.0	6.0	78.0	100.0	0.0	75.0	98.0	2.0
Cumulative	**57.0%**	**96.0%**	**4.0%**	**59.0%**	**97.0%**	**3.0%**	**59.0%**	**94.0%**	**6.0%**	**59.0%**	**96.0%**	**4.0%**
Calibration Standard	**46.7%**	**93.3%**		**37.5%[A]**	**87.5%[a]**		**57.0%**	**100.0%**				
				44.4%[B]	**88.9%[b]**							

[a] CalTPA Instructional Cycle 1
[b] CalTPA Instructional Cycle 2

Results

Using the ratings from the Common Rubric and TPA Model Rubrics, we conducted a series of analyses to examine the comparability of the TPA models. As documented previously, the Common Rubric measures six TPE-elements across all of the TPA models. Table 7.5 maps out how the individual TPA Model Rubrics align, according to TPA model documentation, with each of the Common Rubric TPE elements.

Table 7.5. Alignment of Common Rubric TPE Elements with TPA Model Rubrics

	CalTPA		FAST		edTPA			
Common Rubric TPE Elements	Cycle 1	Cycle 2	Site Visit Project	Teaching Sample Project	Task 1	Task 2	Task 3	Task 4[a]
1.8	R6	R5	IMP	IDM	R5	R8, R10		
3.1	R1, R2, R3, R4	R1	PL, IMP	LO, DfI	R1			
3.2	R1, R2, R3, R4, R8	R1, R2	PL, IMP, REF	LO, DfI, IDM	R2, R3, R4	R10		
3.5	R1, R2, R3, R4	R2, R3	PL, IMP		R4		R14	
5.1		R2, R5, R7		AP	R5		R11	
6.1	R7, R8	R9	REF	R&SE		R10	R15	

Note. R stands for Rubric; FAST acronyms: PL = Planning, IMP = Implementation, REF = Reflection, SiC = Students in Context, LO = Learning Outcomes, AP = Assessment Plan, DfI = Design for Instruction, IDM = Instructional Decision Making, ASL = Analysis of Student Learning, R&SE = Reflection & Self Evaluation.

[a]Task 3 and Task 4 are both Assessment Tasks. During the April 2018 Content Validity Workshop, experts linked TPE elements to the Planning, Instructing, and Assessment Tasks (i.e., Tasks 1 – 3). We might expect linkages to be similar for the second Assessment Task (i.e., Task 4). However, this was not verified and thus no linkages are made. This is discussed in the Limitations section at the end of the report.

As mentioned previously, three additional TPE elements were common to CalTPA and FAST (i.e. TPEs 2.5, 4.1, and 5.2), but not to edTPA. Analysis of CalTPA and FAST using all nine TPE-elements in a Common Rubric resulted in similar conclusions to those observed using the six TPE-elements in a Common Rubric. In other words, including three additional TPE elements in the Common Rubric did not improve the strength of the findings for CalTPA and FAST. Thus, the results in this chapter are based on using the six TPE-elements in the Common Rubric, which are common to all three models. First, we examined Pearson correlations among the Common Rubric total score and Model Rubric total score to determine the extent to which they measure a common domain. Second, we modeled a simple linear regression of the Common Rubric total score on each Model Rubric total score to predict cut scores on the Common Rubric range using each TPA model's cut score. These analyses provided insight as to whether the candidates would pass and fail at similar levels among the three TPAs. Third, using the model-predicted Common Rubric cut scores, we conducted classification consistency analyses to see the degree to which the Common Rubric and each Model Rubric classified candidates the same. For both edTPA and FAST, a single outlier observation was removed. In each instance, the outlier represented an extremely low level of performance and, thus, was removed so as not to have an undue influence on the results.[37]

[37] In each instance, the candidate performed poorly on both sets of rubrics—the Model Rubric and the Common Rubric.

Correlation Analyses

Results for the correlations between the Common Rubric total score and Model Rubric total score are listed in Table 7.6. Moderately strong to strong positive correlations (0.75 for edTPA, 0.41 for CalTPA, and 0.46 for FAST) were observed between the Common Rubric total score and each TPA Model Rubric total score. Additional analyses examined the correlations between each individual Common Rubric TPE-element and TPA Model Rubric mapped to those same TPE elements by the model sponsor (see Appendix 7.B). While the correlation tables depicted in Appendix 7.B do not directly inform the comparability of the TPA models, they may be of interest in understanding the degree to which each of the underlying rubrics from the Common Rubric and the TPA Model Rubrics are correlated with one another.

Table 7.6. Correlations between Common Rubric Scores and TPA Model Rubric Scores

	edTPA	CalTPA	FAST
	.75***	.41*	.46*

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; $n = 29$ for edTPA and FAST; $n = 30$ for CalTPA.

Cut Score Analyses

The cut score analyses were performed by first fitting a simple linear regression of the Common Rubric total score on the Model Rubric total score, and then predicting the cut score and a 95% confidence interval for that prediction on the Common Rubric range using the TPA model cut score.[38] Results for the cut score analyses are presented in Table 7.7 and Figure 7.1. The results in Table 7.7 show the predicted cut score as well as a 95% confidence interval for that prediction. Each graph in Figure 7.1 illustrates these outcomes with the dashed green line indicating the predicted cut score on the Common Rubric range after fitting a linear regression between the total scores and inputting the cut score value for each TPA. The red dashed lines represent the 95% confidence interval around the prediction. Overlapping confidence intervals between the TPA model graphs would suggest that the models are likely to classify candidates similarly. The results indicate that the models' confidence interval ranges do in fact overlap. While the FAST predicted cut score is lower than the predicted cut score for edTPA or CalTPA, the upper range of the FAST cut score confidence interval (21.32) exceeds the lower range of the edTPA and CalTPA cut scores (20.77 and 21.07, respectively), suggesting that the TPA models would classify candidates comparably based on these common TPE elements.

Table 7.7. TPA Model Predicted Cut Scores on Common Rubric Range

	Predicted Cut Score	Lower 95% Confidence Interval	Upper 95% Confidence Interval
edTPA	21.66	20.77	22.55
CalTPA	22.48	21.07	23.89
FAST	19.49	17.66	21.32

Note. $n = 29$ for edTPA and FAST; $n = 30$ for CalTPA.

[38] edTPA requires a score of at least 49 (out of 90); CalTPA sets passing for Cycle 1 to be a score of at least 19 (out of 40) with only one score of "1" and for Cycle 2 to be a score of at least 21 (out of 45) with only one score of "1"; FAST requires at least a score of 2 (out of 4) on all rubrics. The Common Rubric cut score was predicted by inserting into the linear regression models a score of 49 for edTPA, a score of 19 and 21 for CalTPA, and a score of 20 for FAST.

Figure 7.1. Predicted Common Rubric cut scores by TPA model.

Blue line represents simple linear regression of Common Rubric total score on TPA model total score; green dashed line represents the predicted value of the cut score; red dashed lines represents the 95% confidence interval for the predicted cut score; dots represent each candidate's total score on the TPA rubric and Common Rubric. *n* = 29 for edTPA and FAST; *n* = 30 for CalTPA.

Classification Consistency Analysis

Using the predicted cut scores from the previous analysis, we classified candidates according to their scores on the Common Rubric; those with Common Rubric scores below the predicted cut score were classified as failing and those above the predicted cut score were classified as passing. These classifications were compared with classifications after applying the classification rules from the TPA models to the TPA model total scores. These results are reported in Table 7.8. Each 2 x 2 classification consistency table is organized such that the top left cell shows those classified as passing by both rubrics, the top right indicates those classified as failing on the Model Rubric but passing on the Common Rubric, the bottom left indicates those classified as failing on the Common Rubric but passing on the Model Rubric, and the bottom right represents those classified as failing on both rubrics. The larger the share of classifications in the top left and bottom right, the more consistent the classifications across the Model Rubric and Common Rubric. The classification consistency rates (summing across the top left and bottom right) are all quite high with edTPA at 86.13%, CalTPA at 70.00%, and FAST at 79.30%. The strong classification consistency results provide further evidence that the models likely classify candidates in similar ways.

Table 7.8. Classification Consistency Analysis by Common Rubric and TPA Model Rubric

	edTPA		CalTPA		FAST	
	Pass%	Fail%	Pass%	Fail%	Pass%	Fail%
Common Rubric Pass	79.23	3.44	63.33	10	75.86	3.44
	(23)	(1)	(19)	(3)	(22)	(1)
Common Rubric Fail	10.34	6.90	20.00	6.66	17.24	3.44
	(3)	(2)	(6)	(2)	(5)	(1)
Total Classification Consistency%	86.13		70.00		79.30	

Note. n = 29 for edTPA and FAST; n = 30 for CalTPA. Bold cells indicate classification agreement between Common Rubric and TPA Model (i.e., sum of passing on both rubrics and failing on both rubrics).

Discussion

As with the other investigations reported in the previous chapters, the purpose of this activity was to investigate the comparability of the three TPA models. However, unlike the other chapters which addressed comparability by examining the internal functioning of each model and qualitative indicators of comparability, this activity used a common criterion measure to directly compare models. Using the findings from the content validity investigation, in which a panel of teacher preparation experts identified the TPE elements for which there is strong evidence from all three models that the full depth and breadth of the element is measured and in substantively the same way, we developed a Common Rubric. A representative sample of portfolios from each model were then scored by trained scorers using this Common Rubric, and results were compared across models.

The findings from the comparability analyses indicate that scores on the Common Rubric were moderately strong to strongly correlated with the scores from each model's rubric. This supports that, despite the unique components and rubrics for each TPA model, all three models are measuring a highly related construct of teaching performance (based on the subset of TPEs that could be reliably compared). The strongest correlation was between the scores from the edTPA Model Rubric and the Common Rubric (r = .75). Correlations for CalTPA and FAST were also

reasonably strong and of similar magnitude (r = .41 and .46, respectively). To further explore the comparability of the scores, we regressed the Common Rubric scores onto the Model Rubric scores to identify a predicted cut score on the Common Rubric for each model. We computed the 95% confidence interval around each predicted cut score. The findings indicate that the models' confidence interval ranges of cut scores overlap for each model. This suggests that the three models would comparably classify candidates as passing or failing. In other words, regardless of which teaching performance assessment a candidate completes, his/her performance is likely to be consistently classified as passing or failing by all three models (again, based on the subset of TPEs that could be reliably compared). Another way to look at this is through a 2x2 contingency table of pass/fail decisions on each rubric. Those results—which we referred to as the classification consistency analysis—show that the great majority of portfolios included in Activity 7 were consistently classified as pass on both the Model Rubric and the Common Rubric and consistently classified as fail on both the Model Rubric and the Common Rubric. Classification consistency was highest for edTPA (86%), followed by FAST and CalTPA (79% and 70% classification consistency, respectively). As expected, this pattern is consistent with the correlation results. Collectively, the findings from these analyses support that the pass/fail outcomes from each model are comparable when compared to a common criterion measure.

Limitations

It's important to note limitations of this activity. First, this activity captured a limited portion of the TPE construct space. There are 6 TPE domains and within those domains there are 45 TPE elements. Of those 45 TPE elements, there were only six that teacher preparation experts identified as being common to all three models—that is, all three models measured those TPE elements in substantively the same way (i.e., required same or similar evidence) and each model was deemed to assess the full depth and breadth of the TPE element. In light of this, the results from this investigation help to inform how comparable the models are with regard to equitably identifying "TPE-ready professionals," *but* based on a small subset of the TPEs. It is important to keep in mind, however, that all models measure just a subset of the TPE elements, which is consistent with *Assessment Design Standard* 1(a), which states, in part, that "*Each task is substantively related to two or more major domains of the TPEs, and, collectively, the tasks and rubrics in the assessment address key aspects of the six major domains in the TPEs*." In other words, there is no requirement to measure some minimum number of TPE elements, nor are there requirements about which of the 45 TPE elements should be measured. As such, the model sponsors have broad flexibility in deciding which TPE elements and how many of them should be assessed by the TPA. Given that the *Assessment Design Standards* afford this flexibility, it is perhaps not surprising that the Common Rubric was based on just six TPE elements. We could have greatly expanded the number of TPE elements identified as "common" across all three models (i.e., from six to 19) if we would have included those TPE elements that the experts identified as measured by each model, but for which the strength of evidence was "moderate" as opposed to "strong." That is, the models assessed key aspects of the TPE element, but not the full depth and breadth of the TPE element (see Tables 2.1-2.6 in Chapter 2). We opted to exclude those common elements with moderate evidence from our Common Rubric because we believed it was important to only compare models on those TPE elements for which each model measured the full depth and breadth of the TPE element. Including TPE elements for which models assessed key aspects, but not the full depth and breadth, would have compromised the efforts to compare models on a level playing field. Thus, a very stringent criterion was adopted for identifying "common" TPE elements.

Another limitation of this study is that it does not control for population differences among the candidates completing edTPA, CalTPA, and FAST portfolios. A stratified, random sample of

portfolio submissions was obtained from each model. This helped to ensure that the sample of portfolios scored against the Common Rubric were representative of each model's full population. However, just because each model's sample of portfolios was representative of its population does not mean that each model's population of candidates are comparable. If a model has a higher performing population of candidates than another model, then that could impact the comparability results. We know from the subgroup comparison analyses in Chapter 6 that edTPA and CalTPA candidates are demographically similar, whereas the FAST candidates differ demographically from these two models on race. More Hispanic candidates than White candidates take FAST. This is one known population difference and there are likely others.

Third, we adopted a rigorous approach to train and calibrate the assessors who scored candidates on the Common Rubric; however, they were not educators who had experience teaching beginning teachers. This differs from each model's assessor qualification requirements in that all models require that assessors have experience teaching beginning teachers. Instead, the scorers were education researchers with 12+ years of scoring expertise, including experience evaluating scorer training and scoring results in the area of teacher certification. In addition, each of the scorers attended the models' scorer training (either in-person or online). One might argue that their training and experience brings objectivity to their scoring. Nonetheless, it may be that experience teaching beginning teachers is a crucial qualification for scoring. If so, these scorers lacked that experience and we have no way of estimating the impact this may have had on their scoring.

Fourth, due to scope constraints, this activity relied on data for the multiple subject credential. Across the three models, the multiple subject credential is the most frequently sought credential. Had we more time and resources we would have included additional credential areas. However, due to scope constraints, we focused on the multiple subject credential and, thus, these findings may not apply to other credential areas.

A limitation unique to FAST is that the HumRRO scorers' rubric ratings were based on observing a 3 to 5-minute video clip of the candidate's instruction, along with the additional evidence requirements submitted by the candidate (e.g., class profile, lesson plan, commentary). However, the FAST official scorers are in the classroom with the candidate and observe and evaluate the entire 20 to 45-minute lesson. Thus, the HumRRO scorers only had access to 3-5 minutes of the candidates' instruction, whereas the FAST scorers had access to the candidates' instruction for the entirety of the lesson. Despite this difference, the correlation between the Common Rubric scores and the official FAST scores were reasonably strong.

Finally, we know from the model sponsors that some changes were made to CalTPA and FAST following operational Year 1 (2018–19) and that those changes will be reflected in operational Year 2 (2019–20). The information gathered for Activity 7 reflects the status of the models as they were implemented in 2018–19, and, thus, do not reflect updates implemented since the 2018–19 administration. While edTPA did not change in 2018–19, the content validity investigation in April 2018 mapped the TPE elements to the Planning, Instruction, and Assessment Task, but not to the additional Assessment Task (Task 4) included in the elementary education handbooks, which were operational in 2018–19. This is a limitation. The comparability findings for all three models may have been even stronger than what we found, if these aspects could have been included.

Conclusion

This final activity of the comparability investigation provides an innovative and informative approach to addressing the question of comparability across TPA models. By scoring the same candidate submissions on a common criterion measure, this activity provides yet another indication of comparability, albeit on a subset of the TPE elements, although, again, it is important to keep in mind that all models measure just a subset of the TPEs. The findings reported here indicate that despite the unique components and rubrics for each TPA model, all three models tap into a common construct of teaching performance and that regardless of which model a candidate completes, all three models are likely to consistently identify a candidate as passing or failing. In other words, the findings from Activity 7 lend additional support to the claim that all three models "equitably identify TPE-ready professionals."

Chapter 8: Summary

A comprehensive, two-year investigation was undertaken to investigate the comparability of the three teaching performance assessments approved in California as a credentialing requirement for beginning teachers. Thus, the objective of this investigation was to compare the three TPAs on key aspects of test design, implementation, scoring, and reporting to create a body of evidence and thereby triangulate—that is, capture from different angles—whether the models are indeed comparable. To accomplish this ambitious objective, we designed and conducted seven activities, each investigating a claim(s) that should be substantiated to support model comparability.

An overall summary of each activity and its results is provided in Table 8.1.

Table 8.1. Summary of Body of Evidence

Claims	Activities Investigating Claims	Overall Conclusions
1. TPAs are sufficiently comparable in their representation of the ADS and in their assessment and weighting of the TPEs.	Activity 1: Evaluation and comparison of Evidence across TPAs for Adherence to ADS[a] Activity 2: Content Validity Comparability Analysis Activity 3: Surveys of Stakeholders	• All TPAs mostly or fully adhere to the ADS and relevant test design *Joint Standards.* (Act.1) • All TPAs exceed the requirement for ADS 1(a) by having all tasks/cycles assessing three or more TPE domains. (Act. 2) • TPE 3 is the domain assessed most thoroughly by all TPAs. (Act. 2) • TPE 6 is the domain assessed least thoroughly by all TPAs. The experts reported that TPE 6 is difficult to measure via performance assessments. Thus, it's important that programs are addressing TPE 6 through means other than the TPA. (Act. 2) • There are some differences in emphasis and measurement of TPE elements across TPAs, with FAST and CalTPA being more comparable and having slightly stronger evidence linkages to TPE elements than edTPA, particularly for TPE domains 2 and 4. (Act. 2) • Overall, candidates and coordinators perceive their model as valid. (Act. 3)
2. Guidance and supports provided by model sponsors to candidates and coordinators is sufficiently clear and detailed to ensure the TPA is implemented as designed and intended.	Activity 3: Surveys of Stakeholders	• Resources (e.g., manual/handbook/guide, website) were perceived as helpful by most candidates and coordinators. The online system for uploading FAST submissions was the only resource identified for improvement. • The majority of candidates (particularly for FAST) and coordinators reported having a clear understanding of their model's requirements (e.g., what to submit as evidence). • These findings coupled with the perceived validity of the TPAs should help to ensure that models are implemented as designed and intended.
3. Scoring rubrics for each TPA are sufficiently clear and detailed to ensure that trained scorers can accurately and consistently score submissions.	Activity 4: Scoring Review	• Overall, all TPAs mostly or fully adhere to the ADS and *Joint Standards* related to rubrics. • The format and structure of the edTPA and CalTPA rubrics are similar; both are analytic with five levels labeled Level 1, 2, 3, 4, 5. • The FAST rubric has four score levels with labels that range from "Does Not Meet Expectations" to "Exceeds Expectations." Each of the 10 rubrics contains 2–3 indicators. Additional guidance on how FAST scorers should combine indicator level ratings to arrive at an overall rating for each rubric may help to further strengthen scorer consistency. • Scoring rubrics for FAST and CalTPA are mapped to TPE elements and included in the candidate manual/assessment guide. edTPA may want to consider making this same information available to its candidates, programs, and assessors. • Few exemplars at the extremes of the scale (i.e., Level 1s and Levels 4 and 5) were observed in scorer training. Additional exemplars of these score levels are recommended for inclusion in scorer training.
4. There is a comparable, comprehensive process to select, train, and establish calibration of the assessors who score submissions.	Activity 4: Scoring Review	• Scoring processes for all TPAs address key aspects of the ADS and *Joint Standards* related to scorer training. • edTPA and CalTPA have stronger procedures to ensure that scorers maintain the calibration attained during training; FAST should implement a scorer consistency check during the scoring window. • Returning scorers should be required to re-calibrate; this was not a requirement for FAST in 2018-19.
5. The standard-setting procedures for each TPA are sufficiently comparable and (continued)	Activity 5: Comparison of Standard Setting across TPAs	• edTPA and CalTPA used procedures (Briefing Book method) that are sufficiently comparable and rigorous to ensure that their passing standards accurately and consistently identify candidates possessing the requisite KSAs.

(continued)

Table 8.1. (Continued)

Claims	Activities Investigating Claims	Overall Conclusions
5. (cont'd) rigorous to ensure that the respective passing standards for each model accurately and consistently identify candidates possessing the requisite KSAs.	Activity 5: (continued)	• FAST used a non-traditional standard setting method whereby teacher preparation staff reviewed rubric descriptors to ensure that Level 2 ("Meets Expectations") descriptors adequately described KSAs of a just sufficiently qualified beginning teacher. Future standard setting activities for FAST should consider including actual candidate submissions, impact data, and consideration of a compensatory scoring model.
6. Each TPA conducts statistical analyses to identify differential effects in relation to candidates' race, ethnicity, language, gender or disability.	Activity 6: Statistical Analysis and Comparison of Score Data	• There was no evidence of substantive differences in pass rates for males and females within TPAs; also, the pattern of pass rates for males and females is similar across TPAs. • There were similarly small differences between mean total scores for males and females for all TPAs. • Racial demographics for edTPA and CalTPA are similar. FAST differs from edTPA and CalTPA in that the majority of FAST candidates are Hispanic, not White. • Pass rates for the various race categories were similar both within and across models for edTPA and CalTPA (FAST excluded from this analysis due to population difference). • Comparisons of mean total scores between race groups showed no notable differences across models, except that Whites tended to have higher mean total scores on FAST than Hispanics, although this did not translate to differences in pass rates. • All models should collect additional data on other demographic variables—notably, language and disability—so that ADS 1(k) can be fully investigated.
7. Score reports (candidate and program) provide similar information about candidate outcomes and include clear guidance on how candidate score information should be used.	Activity 4: Scoring Review	• All TPAs mostly or fully adhere to the ADS and *Joint Standards* relevant to score reports. • CalTPA and edTPA score reports are more similar to one another than either is to FAST score reports. For example, FAST score reports only provide rubric level scores, not total scores or passing status (although FAST candidates know that they must obtain a '2' on all 10 rubrics to pass). • CalTPA and edTPA score reports include guidance that scores are used to compare candidates' performance (knowledge and skills) to the requirements set by the Commission (their state). • No TPAs include guidance on their score reports that scores should be used in conjunction with other measures to determine a candidate's readiness for beginning teaching, although all models include this information in other supporting material. Ideally, this information would appear directly on score reports.
8. Rubrics and score reports provide diagnostic information on candidates and on programs such that the strengths and weaknesses of each can be identified.	Activity 4: Scoring Review	• Rubric level scores reported on score reports for all models are, in and of themselves, diagnostic, although only CalTPA score reports include specific guidance on score reports that rubric level scores "may help you identify your relative strengths and areas of improvement." However, FAST and edTPA include similar language in other documents. • CalTPA and edTPA model sponsors provide programs access to an online platform to analyze program level results. FAST, being a local program, already has this information.
Ultimate Objective: all models equitably identify TPE-ready professionals.	Activity 7: Comparison of TPA Models to a Common Criterion	• When a representative sample of multiple subject portfolios was scored on a Common Rubric (measuring a subset of TPE elements identified from Activity 2 for which there is strong evidence that all models measure the full depth and breadth), the pass/fail outcomes on the Common Rubric were consistent with the pass/fail outcomes on each model's rubric. This suggests that regardless of which TPA a candidate completes, his/her performance is likely to be consistently classified as passing or failing by all TPAs (based on these elements).

[a]Activity 1 serves as an overarching investigation of all eight claims, although it most directly addresses Claim 1.

Practical Implications

The primary practical implication of this investigation is that it provides empirical evidence to support the Commission's decision to approve multiple TPA models as a credentialing requirement for beginning teachers. Again, this is not to say that the models are equal, but rather that all models are likely to equitably identify teacher candidates who are "ready"—that is, possess the KSAs required for beginning teaching. The findings from this investigation do point out some potential threats to the comparability of the TPAs, which the model sponsors are encouraged to address. Doing so will further strengthen model comparability, as well as the quality and rigor of the TPA model. If the Commission is concerned about differences across the models in the representation of the TPE *elements* assessed, then to further strengthen model comparability the Commission might provide the model sponsors with guidance at the level of TPE *elements*, rather than TPEs overall. This could be done through a modification to the *Assessment Design Standards*. This investigation shows that the ADS have provided a strong blueprint for the models to follow and that the model sponsors are closely adhering to the ADS. This suggests that any changes the Commission might make to the ADS are likely to be enacted by the model sponsors.

Future Research

This body of research demonstrates a comprehensive investigation of TPA model comparability. Nonetheless, additional research is recommended to further support the validity argument for model comparability. Validity arguments are not static, rather they are dynamic and are strongest when supported by ongoing research to support continuous improvement. Suggestions for future research include an expansion or elaboration upon the studies conducted herein, and new avenues of research. Some of the ways the studies conducted herein could be expanded or elaborated upon are outlined below.

- Conduct another content validity investigation (Activity 2) but expand upon it by having teacher preparation experts identify which aspects of each TPE element are assessed by each model. In the current effort, a strong evidence linkage indicated that the model assessed the full depth and breadth (i.e., all aspects) of the TPE element. Thus, these were the TPE elements included in the Common Rubric in Activity 7. A moderate evidence linkage indicated that the model assessed key aspects of the TPE element, but not the full depth and breadth. Because we wanted to ensure that all models would be compared on a level playing field in Activity 7, only TPE elements for which there was "strong evidence" across all three models were included in the common rubric. However, if we identified the key aspects of the TPE elements that received "moderate evidence" ratings, then we could identify additional teaching performance expectations that are common to each model.
- The above bullet point could be further extended by creating a new common rubric that more fully addresses the construct space, and then updating Activity 7 using this more robust common rubric.
- Activity 6 (investigation of score patterns for subgroups) could be conducted for other credential areas beyond the multiple subject credential. Also, Activity 6 could be expanded upon by investigating subgroup differences in score patterns for language and disability, assuming the models capture this demographic information in their score data. If multiple years of data were combined, then this would help to circumvent concerns regarding small samples.

- When/if notable changes are made to a model(s), any number of the seven activities could be repeated to capture these updates.

There are also new avenues of research that could supplement this existing body of research. Some new areas of research might include:

- A longitudinal, predictive validity study in which candidates' scores on their TPA are correlated with a measure of their teaching performance (e.g., their performance evaluation from their first year of teaching). Such a study would address an important gap in the validity argument for the TPA models—i.e., it would provide empirical evidence that the models are indeed predictive of the KSAs necessary for beginning teachers.
- A convergent validity study in which candidates' scores on the TPA are correlated with another assessment of teaching. Another assessment to potentially explore is the California Basic Educational Skills Test (CBEST), which measures candidates content knowledge in the areas of reading, mathematics and writing, and for which scores are readily available to the Commission. One might expect that candidates' scores on their single subject mathematics portfolio, for example, may correlate more strongly with their scores on the mathematics portion of CBEST (convergent evidence) than with their scores on the reading portion of CBEST (discriminant evidence)—i.e., teachers must know the content areas they teach. Such information would help to support the construct validity evidence for the TPAs and for the CBEST alike.

The research listed above would not only further support the validity argument for model comparability but would also further strengthen the validity evidence for any given model.

Final Conclusion

As with all research studies, these seven activities are not without their limitations, and those were described within the body of the report. Nonetheless, this investigation paints a rich picture of comparability. Certainly, there are differences across the TPA models. In many cases those differences do not pose threats to the veracity of the claims and the differences are in line with the Commission's expectations—as evidenced by the fact that the Commission's *Assessment Design Standards* allow for flexibility in how each model assesses the TPEs. However, some of the identified differences may pose threats to the veracity of the claims and, ultimately, to the equitable identification of "TPE-ready professionals." In this regard, this report should serve a formative purpose for the model sponsors so that they can address potential threats to model comparability.

The Commission should be commended for undertaking a comprehensive investigation of the comparability of the TPA models. Not only does this investigation bolster support for the claim that the TPA models are comparable, it also strengthens the validity evidence for each of the models. As such, the Commission can be assured that there is compelling validity evidence to support each of the models they have approved. As one of the TAC members commented, this investigation may serve as a useful roadmap for other states and/or credentialing organizations that are considering approving multiple performance assessments for credentialing decisions.

References

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Commission on Teaching Credentialing (2015). California Teaching Performance Assessment Design Standards. Retrieved from: https://www.ctc.ca.gov/docs/default-source/educator-prep/tpa-files/tpa-assessment-design-standards.pdf?sfvrsn=2e393153_0

Haertel, E. H. (2005). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21,* 16-22. doi:10.1111/j.1745-3992.2002.tb00081.x

Haertel, E. H. (2008). Standard setting. In K.E. Ryan and L.A. Shepard (Eds.), The future of test-based educational accountability (pp139-154). Routledge.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24(4),* 355 – 366.

Hambleton, R. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Erlbaum.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). https://humrro.sharepoint.com/sites/Resource-Center/Shared Documents/Forms/AllItems.aspx?viewid=19a6c4bb%2D35f0%2D4527%2D80a8%2D5c332ed506b6&id=%2Fsites%2FResource%2DCenter%2FShared%20Documents%2FBrownbags%2FBB%20ScheduleWestport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), pp. 1-73.

Renter, D. G.; Higgins, J. J.; and Sargeant, J. M. (2000). Performance of the Exact and Chi-Square Tests on Sparse Contingency Tables, Conference on Applied Statistics in Agriculture. https://doi.org/10.4148/2475-7772.1253

Sinclair, A. L. (2017). *An investigation of the comparability of commission-approved teaching performance assessment models: Project implementation plan*. Alexandria, VA: Human Resources Research Organization.

Sinclair, A. L. & Thacker, A. (2018). Content Validity Comparability Analysis. In Sinclair, A. L. & Thacker, A. (Eds.), An investigation of the comparability of commission-approved teaching performance assessment models: Year 1 Preliminary Report (2018 No. 047). Alexandria, VA: Human Resources Research Organization.