edTPA®

EDUCATIVE ASSESSMENT & MEANINGFUL SUPPORT

2019 edTPA Administrative Report

Released June 2021

Preface and Acknowledgments

edTPA is a performance assessment for initial licensure teacher candidates, which was developed and field-tested beginning in 2009 and used operationally since September 2013. This report presents analyses based on teacher candidate performance from **January 1 to December 31, 2019,** and complements prior reports that have reviewed the development of the assessment, as previously described in detail in the 2013 edTPA Field Test Summary Report and the 2014 through 2018 edTPA Annual Administrative Reports. As such, this 2019 report takes a streamlined approach and does not re-state prior results; rather it refers readers to existing descriptions of important program features, while providing updated 2019 Program Data. The Field Test Report and all annual administrative reports are located in the edTPA Resource Library.

This administrative report was authored by Raymond L. Pecheone, Executive Director, Stanford Center for Assessment, Learning and Equity (SCALE); Nicole Merino, Director, Teacher Performance Assessment, SCALE; Heather Klesch, Director, Educator Solutions for Licensing and Learning, Evaluation Systems group of Pearson; Jon Twing, Senior Vice President, Pearson Global Assessment Services; and Lori Kroeger, Senior Educational Consultant, Evaluation Systems group of Pearson.

SCALE is the sole developer of edTPA, and Stanford University is the exclusive owner of edTPA. The university has a licensing agreement with the Evaluation Systems group of Pearson to provide operational support for the national administration of edTPA.

As authors of this report, we welcome all comments regarding this report and its data and will carefully consider such comments as we continue to research, enhance, and improve edTPA as a support and assessment system.

Table of Contents

Executive Summary

The Stanford Center for Assessment, Learning and Equity (SCALE), the American Association of Colleges of Teacher Education (AACTE), and the Evaluation Systems group of Pearson are pleased to release the 2019 Administrative Report. This report presents candidate performance data from over 49,100 candidates who participated in edTPA during the sixth full operational year (January 1 to December 31, 2019), associated analyses affirming reliability of scoring, and validity evidence supporting its intended use as a measure of readiness to teach and a metric used to inform program approval or accreditation. The analyses and results presented here are intended to meet the technical standards for licensure assessments set forth by AERA, APA, & NCME (2014).

SCALE and AACTE commend the over 900 educator preparation programs in 41 states and the District of Columbia that have contributed to the development and field testing of edTPA and its use since 2009. We also commend the teaching candidates who have engaged with edTPA during the development stages, and since the operational launch in September 2013, as a reflective experience that demonstrates the knowledge, skills, and abilities embedded in their real teaching with real students in real classrooms across the country. Moreover, edTPA was purposefully designed to reflect the job-related teaching tasks that are represented in the National Board for Professional Teaching Standards (NBPTS) as it pertains to the skills and competencies attained as part of teacher preparation.

Developed by subject-specific faculty design teams and staff at SCALE, with ongoing input from hundreds of teachers and teacher educators from across the country, **edTPA is the first nationally available, educator-designed support and assessment system for initial teacher licensure**. It provides a measure of teacher candidates' readiness to teach that can inform licensure, accreditation decisions, and program completion. Importantly, edTPA is an educative assessment that supports candidate learning and preparation program renewal.

edTPA Design

edTPA is a subject-specific performance assessment that evaluates a common set of teaching principles and teaching behaviors as well as pedagogical strategies that are focused on specific content learning outcomes for P–12 students (birth–21 in special education settings). SCALE's extensive Review of Research on Teacher Education provides the conceptual and empirical rationale for edTPA's three-task design and the rubrics' representation of initial competencies needed to be ready to teach. The assessment systematically examines a cycle of teaching aimed at subject-specific student learning goals, using authentic evidence derived from candidates' practice in their student teaching or internship placement. A cycle of teaching, captured by the three tasks that compose an edTPA portfolio, includes:

1. planning,
2. instruction, and
3. assessment of student learning.

Authentic and job-related evidence includes lesson plans, instructional materials, student assignments and assessments, feedback on student work, and unedited video recordings of instruction. Also assessed through the three tasks are candidates' abilities to develop their students' academic language and to justify and analyze their own teaching practices.

All 28 edTPA handbooks share approximately 80% of their design, assessing pedagogical constructs that underlie the integrated cycle of planning, instruction, and assessment. This shared architecture and common underlying pedagogical constructs, like those represented through the NBPTS portfolio, lay the foundation for the single standard of pedagogical performance with an emphasis on content specificity. The other 20% features key subject-specific components of teaching and learning drawn from the content standards for student learning and pedagogical standards of national organizations. For example, consistent with the National Council of Teachers of Mathematics standards, the elementary, middle childhood,

and secondary mathematics versions of edTPA require candidates to demonstrate subject-specific, grade-level-appropriate pedagogy in mathematics. The Secondary Mathematics edTPA handbook requires that candidates support students' development of conceptual understanding, procedural fluency, problem solving/reasoning skills, and precise communication skills for a standards-based topic; support identified academic language demands; and include opportunities for students to demonstrate development of mathematics concepts, reasoning skills, and mathematical precision.

edTPA's Educative Purpose – A Support and Assessment System

Unlike typical licensure assessments external to programs, edTPA is intended to be embedded in a teacher preparation program and to be "educative" for candidates, faculty, and programs. Candidates deepen their understanding of teaching while preparing for edTPA and reflecting on their strengths and challenges as they move forward into the profession. For faculty and programs, the various edTPA resources and candidate, program, and campus results can be used to determine areas of program strength and curricular renewal (see Whittaker, Pecheone, & Stansbury, 2018; Pecheone & Whittaker, 2016). In addition, the professional growth plan resource uses edTPA results and other evidence of teaching to inform candidates' goal setting for induction and the early years of teaching.

Since edTPA launched its first "online community" in 2011, membership has grown to over 10,700 faculty from over 928 teacher preparation programs. The website (edtpa.AACTE.org) also includes publicly available materials for various stakeholders. In addition to the website, edTPA offers a National Academy of experienced consultants available to provide professional development to new users and to network in a learning community across the country. Lastly, programs using edTPA are provided with a variety of tools and reporting formats to access, analyze, and make decisions about their own candidate performance data, as well as state and national summary reports.

Scorer Training, Monitoring, and Reliability of Scores

Educators play a critical role in the scoring of edTPA. Over 3,100 qualified teachers and teacher educators now serve as scorer trainers, supervisors, or scorers. Scorers must be current or recently retired (within the past 5 years) P–12 teachers or administrators or educator preparation program (EPP) faculty, field supervisors, or program administrators with significant pedagogical content knowledge in the field in which they score. Additionally they must have experience working as instructors of teacher candidates or mentors for novice teachers (e.g., NBPTS teachers, hosting or mentoring of teacher candiates). Furthermore, scorers are matched to both their subject of expertise and grade span for the discipline in which they will score. Approximately 50% of scorers are from each group (i.e., P–12 or EPP).

Before becoming an official edTPA scorer, educators must go through an extensive scorer training curriculum developed by SCALE and meet qualification standards demonstrated by scoring consistently and accurately. Once scorers qualify and score operationally, they are systematically monitored during the scoring process (through quality monitoring processes such as backreading, validity/calibration portfolios, and requalification exercises) to ensure that they continue to score reliably, including adjudication and resolution measures.

Scorer reliability was evaluated using several different statistical tests. In a random sample of 4,200 portfolios double-scored independently by two scorers, the scorers assigned either the same or adjacent scores (exact + adjacent agreement) in approximately 96% of all cases. More than 30% of edTPA portfolios are double scored, including an independent 10% random "reliability" selection. In addition to this 10% selected for random double scoring, portfolios within a defined range above and below the state-specific cut score (or national cut score when no state cut score is in place) are also double scored. This accounts for approximately 20% of double scored portfolios. In all double scoring cases, the final score is based on at least two scorers who agree. Adjudication and resolution are also implemented as part of double scoring. If Scorer 1 and Scorer 2 are more than 1 score point apart on any rubric, the portfolio is resolved by a scoring supervisor. If the total scores for all 15 rubrics from Scorer 1 and Scorer 2 are on opposite sides of

the cut score, the portfolio is adjudicated by a scoring supervisor. If scorers have five or more adjacent rubric scores (scores that are one point apart) out of 15 rubrics, the portfolio is scored by a scoring supervisor. These double-scoring procedures support the final scores assigned to edTPA candidates.

Reliability and Validity

As in previous reports, in the 2019 edTPA Administrative Report we present validity evidence about internal structure, including reliability coefficients. These results are intended to meet the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) technical standards for licensure assessments of this type and support the use of edTPA scores as a reliable, consistent estimate of a prospective teacher's readiness to teach.

This report includes evidence about consistency in scoring. One component of this report examines inter-rater consistency, including agreement rates, which were evaluated using the kappa-N statistic and the interclass correlation. Kappa-N reflects agreement attributable to scorers who assign scores to categories (even at random), but is inline with expected marginal base rates (Hsu & Field, 2003). Given the training and qualification of edTPA scorers, the expectation is a large base rate agreement and, therefore kappa-N was chosen as the agreement statistic. Statistics such as kappa, including kappa-N, are intended to be used with other methods of consistency included in this report.

Following the validity guidelines presented in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), previous reports define the constructs assessed by edTPA and present evidence that examines its use and interpretations. The validity sections review sources of validity evidence for edTPA. These include the empirical research and theory on which the development was based; the design process and content validity as well as job relatedness; development to ensure that the assessment represents the skills, knowledge, and abilities that represent a candidate's readiness to teach; and evidence based on content and internal structure. We present results from Confirmatory Factor Analyses (CFA) and a polytomous item response theory (IRT) model that provide empirical support for the edTPA constructs of planning, instruction, and assessment. The current report affirms our previous findings.

Candidate Performance

This report presents performance data from 49,123 first-attempt officially scored edTPA submissions including average scores and distributions overall by task and by rubric for all 15-rubric handbooks, including the first 15 rubrics of the 18-rubric handbooks. The Elementary Education: Mathematics with Literacy Task 4 Handbook was made available to the field in 2018; as such, this report provides the intial analyses whereby this handbook could have been included. The Educational Technology Specialist and Library Specialist Handbooks have been excluded due to low usage (N of 7 and N of 3, respectively). The total score, computed as an aggregation of scores on a 5-point scale across 15 rubrics, ranges from 15 to 75 total points. The average edTPA score across 49,123 portfolios from fields with 15-rubric handbooks (including the first 15 rubrics of Elementary Education: Literacy with Mathematics Task 4 and Elementary Education: Mathematics with Literacy Task 4 handbooks) was 44.16 with a total standard deviation of 6.1 and a standard error of the mean of .03. Performance by task is an aggregation of scores on the 5 rubrics that make up each of the three tasks; these range from 5 to 25 points for each task.

Scores across content fields were examined overall, as well as disaggregated based on state-wide policy regarding consequential edTPA use—that is, whether or not the results of edTPA are used to make consequential decisions about candidates or programs. Based on the national recommended professional performance standard of 42 (to date no state has a cut score of 42), the modeled pass rate for all candidates who submitted an edTPA portfolio in 2019 was 72.0% across all states and 75.0% in states using the assessment consequentially. Note that cut scores vary by state as do passing rates, and current state cut scores range from 35 to 41. See details in the body of the report for pass rates by cut score.

Due to large differences in sample size, populations represented within the sample, and small numbers of total submissions in certain fields, interpretations and comparisons across fields should be approached

with caution and should not be generalized across the entire population.

When submitting an edTPA portfolio for official scoring, the candidate is asked to provide demographic information in several categories: gender, ethnicity, teaching placement context (including school location [city, rural, suburb, or town]), length of placement prior to completing edTPA, student characteristics (% English learner, % special education eligibility, % free/reduced lunch), program type (traditional or alternative route), education level, and primary language. Portfolios submitted in states that have policy for consequential use of edTPA were used to examine performance by these demographic categories (N = 41,822). These analyses revealed that all demographic variables taken together explained less than 3% of the total variance in edTPA scores. We present descriptive statistics by demographic categories later in this report. edTPA is committed to providing equal opportunity for all teacher candidates and continues to explore research in this area as well as monitor candidate performance, scorer training, assessment design, and implementation for any potential sources of differential impact.

Conclusion

As with the Field Test data (2013) and those of the previous administrative reports for 2014 through 2018, data presented here are intended to meet the technical standards of APA, AERA, and NCME (2014) and support the use of edTPA as one metric of a multiple-measure system in the decision to grant an initial license to teacher candidates as well as to inform state and national accreditation. The reporting of performance of all candidates who submitted edTPA portfolios in 2019 is presented for all content fields, except those noted above, and informs the use of edTPA across states.

As is the case with NBPTS, educative use of a performance-based assessment is more than a testing exercise completed by a candidate. edTPA's design and emphasis on support for implementation mirrors the NBPTS use of professional networks of experienced users to assist others as they prepare for the assessment. The opportunities for educator preparation program faculty and their P–12 partners to engage with edTPA are instrumental to its power as an educative tool. The extensive and growing library of over 180 resources developed by SCALE, the National Academy of consultants, and state infrastructures of learning communities for faculty and program leaders promotes edTPA as a tool for candidate and program learning. As candidates are provided with support to develop and practice the constructs embedded in edTPA throughout their programs and reflect on their edTPA experience with faculty and P–12 partners, they are more likely to internalize the cycle of effective teaching (planning, instruction, and assessment) as a way of engaging in their practice—as a way of focusing on knowing their students and student learning and helping sustain them in the profession well beyond their early years in the classroom.

External Links to Program Features

The 2019 Administrative Report is a streamlined version similar to the structure of the 2018 report. We present the most up-to-date information about program participation and resources, followed by the findings for validity and reliability, and candidate performance. Links to previously published reports and resources are provided for the following topics:

General
National Academy

edTPA Administrative Reports
edTPA Administrative Reports
States Participating in edTPA and State Policy Summary

Scoring, reported in the 2016 edTPA Administrative Report
Scorer Training
Scoring Model
Standard Setting – National and State

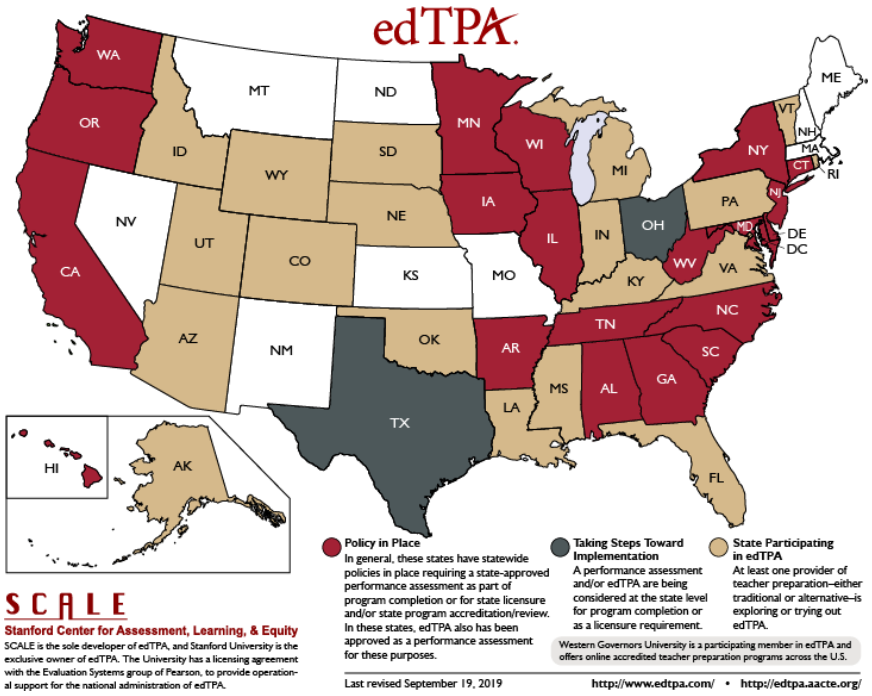
Policies
Candidate Submissions, Originality, Score Confirmation, and Retake Policies

Validity Evidence, reported in the 2016 edTPA Administrative Report
Content Validity and Job Analysis
Construct Validity
Consequential Validity
Concurrent Validity
Predictive Validity

Introduction

States Participating in edTPA

The map below shows the states participating in edTPA as of September 2019. The map indicates states with edTPA implementation policy in place and those expecting to have an implementation policy in place soon. Visit the edTPA website for up-to-date information on states and their policies.



Role of the Partners

edTPA was created for the profession by the profession, with input from a national design team and subject-specific design teams made up of teachers and teacher educators across the country in a process led by SCALE. edTPA is supported by AACTE and the Evaluation Systems group of Pearson. Each partner plays an important role in the design, development, and ongoing operations of edTPA as a nationally available performance assessment of teaching.

Stanford University faculty and staff at SCALE developed edTPA and are the sole authors. Stanford University exclusively owns all of the intellectual property rights and trademark for edTPA. SCALE is responsible for all edTPA development including candidate handbooks, scoring rubrics and the scorer training design, scorer training curriculum, and materials as well as support materials for programs, faculty, and candidates. SCALE also recruits, reviews, trains, and endorses National Academy consultants who act as support providers within the edTPA community.

AACTE partners with edTPA to support development and implementation and disseminates resources via edTPA.AACTE.org so that teacher preparation programs and faculty using edTPA have the materials they need to support teacher candidates. AACTE also supports the deployment of National Academy consultants via the website and an online community forum for networking and program assistance.

Stanford University/SCALE engaged Evaluation Systems, a group of Pearson, as an operational partner in March 2011 to make edTPA available to a national educational audience. As the operational partner, Evaluation Systems provides the management system required for multistate use of edTPA, including the infrastructure that facilitates administration of the assessment for candidate registration, submission, scoring, quality assurance, and reporting of results from both national and regional scoring. Evaluation Systems also recruits scorers (who consist of educators from P–12 schools and educator preparation programs), manages the scoring pool, monitors scoring quality, and provides a delivery platform for the SCALE-developed scorer training curriculum. Evaluation Systems collects and records the scores generated by qualified scorers.

edTPA as a Support and Assessment System

Unlike typical licensure assessments external to programs, edTPA is intended to be embedded in a teacher preparation program and to be "educative" for candidates, faculty, and programs. Candidates deepen their understanding of teaching through use of formative resources and materials while preparing for edTPA, and the score reports provide feedback on candidates' strengths and challenges as they move forward into their first years of teaching. For faculty and programs, the various edTPA resources and candidate, program, and campus results can be used to identify areas of program strength and determine areas for curricular renewal (see Pecheone & Whittaker, 2016; Whittaker, Pecheone, and Stansbury, 2018).

Summary of Resources

Since edTPA launched its first "online community" in 2011, membership has grown to 10,707 faculty from over 928 teacher preparation programs who have access to more than 180 resources including candidate handbooks, rubrics, templates, support guides for candidates, local evaluation protocols, retake guidelines, guidelines for supervising teachers, and webinars addressing edTPA constructs such as Academic Language. The website, edTPA.AACTE.org, also includes publicly available materials for various stakeholders (for example, video and webinar explanations of edTPA and its benefits). **Materials in the resource library have been downloaded over 1.7 million times.**

The most commonly downloaded resources include:

Resource	Number of Downloads
Understanding Rubric Level Progressions 2017–2018 - Complete Collection	147,168
Academic Language Summary Handouts 2017–2018 - Complete Collection	121,798
edTPA Administrative Reports	106,063
State Policies and EPP Implementation Chart	102,669
edTPA Handouts to Share with Stakeholders	74,640
edTPA FAQ	52,210
edTPA Guidance to Supervising Teachers	50,173
Making Good Choices - A Support Guide for edTPA Candidates - Complete Collection	26,127
Elementary Education: Literacy with Mathematics Task 4 Handbook	25,695
Review of Research on Teacher Education: edTPA Task Dimensions and Rubric Constructs	24,584

Reliability and Validity Evidence

Inter-Rater Agreement

The table below shows inter-rater agreement for the 2019 edTPA administration cycle. The table shows agreement rates for each rubric as well as for judgments overall. Inter-rater agreement (IRA) measures to what extent multiple raters provide ratings of items or performance tasks consistently. The check of inter-rater agreement is part of the general quality control for a scoring process, and it requires a process that randomly assigns portfolios to be read by two scorers, independently. It is customary to summarize IRA for three levels of granularity (Chodorow & Burnstein, 2004; Powers, 2000; Stemler & Tsai, 2008), such as:

- Exact agreement – proportion of cases in which the first and second scores match exactly;
- Adjacent agreement – proportion of cases in which the first and second scores are apart by one score point, in absolute value; and
- Exact + Adjacent agreement – proportion of cases in which the pairs of scores are +/- 1 score point apart from each other.

The data set included 4,200 complete submissions from 13-rubric, 15-rubric, and 18-rubric handbooks that were scored independently by two scorers as part of the random sample of double-scored portfolios for the 2019 administration cycle. (Note: 13-rubric handbooks do not have rubrics 4 and 14.) Across all 15 rubrics, independent scorers assigned the same score (exact agreement) in 56.7% of all cases. Scorers assigned scores that were one point apart (adjacent agreement) in approximately 39.3% of all instances. When combining exact and adjacent agreement into an exact + adjacent agreement, scorers assigned scores that were the same or +/- 1 point apart in approximately 96.0% of all cases. These exact and adjacent agreement rates are consistent with past edTPA operational years, and with that of other performance assessments, such as the NBPTS.

The kappa-N provides chance-corrected agreement, or inter-rater agreement measures that result from removing exact + adjacent agreement that may have occurred randomly (Brennan & Prediger, 1981). Chance-corrected agreement ranges from zero to one. There are no widely accepted guidelines for what constitutes an adequate value of the coefficients, although higher values represent greater levels of agreement. The table below shows kappa-N for both exact agreement and exact + adjacent agreement. The kappa-N exact agreement ranged from 0.368 (rubric 12) to 0.709 (rubric 6), with an average value of 0.459; the kappa-N exact + adjacent agreement ranged from 0.857 (rubric 12) to 0.982 (rubric 6), with an average value of 0.915. This outcome corroborates that scorers tend to assign scores within +/- 1 and rarely assign scores that differ by more than 1 point.

RELIABILITY AND VALIDITY EVIDENCE
Inter-Rater Agreement

Task	Rubric	Inter-Rater Agreement				
		Exact	Adjacent	Exact +Adj.	Kappa-N Exact	Kappa-N Adj+Exact
Task 1: Planning	Rubric 01	0.597	0.372	0.969	0.496	0.936
	Rubric 02	0.536	0.400	0.936	0.420	0.867
	Rubric 03	0.550	0.414	0.964	0.438	0.924
	Rubric 04	0.560	0.402	0.961	0.450	0.920
	Rubric 05	0.561	0.393	0.954	0.451	0.903
Task 2: Instruction	Rubric 06	0.767	0.224	0.991	0.709	0.982
	Rubric 07	0.595	0.379	0.974	0.494	0.945
	Rubric 08	0.577	0.394	0.972	0.471	0.941
	Rubric 09	0.561	0.402	0.963	0.451	0.922
	Rubric 10	0.567	0.403	0.970	0.459	0.938
Task 3: Assessment	Rubric 11	0.525	0.412	0.938	0.406	0.870
	Rubric 12	0.494	0.437	0.932	0.368	0.857
	Rubric 13	0.505	0.436	0.941	0.381	0.877
	Rubric 14	0.563	0.406	0.970	0.454	0.937
	Rubric 15	0.544	0.414	0.958	0.430	0.912
Overall Average		0.567	0.393	0.959	0.459	0.915

Internal Consistency

The intraclass correlation coefficient (ICC) is a reliability estimate that provides information about both the degree of correlation and agreement between measurements. For edTPA, SAS GLM was used to obtain information necessary for calculating the ICCs following the procedures of McGraw & Wong (1996). The ICC is an estimate of interrater reliability and shows the variation between two raters across candidates, and is provided for edTPA handbooks with 35 or more double-scored reliability submissions. The table also provides the 95% confidence interval (CI), quantifying the stability of the estimate. The ICC ratings across subject areas ranged from .46 (Middle Childhood Math) to .78 (Middle Childhood History/Social Studies). For interpretations, readers are directed to Koo & Li (2016, pg. 161). For the purposes of the table, these ICCs were obtained assuming one-way random effects, absolute agreement, using a single rater/measure.

Note: The submissions used in this analysis had scores reported in 2019, were part of the reliability double scoring sample, and included the 15 national rubrics. Language fields of study were removed as were any fields without at least 30 pairs of candidate scores.

Intraclass Correlation Coefficient (ICC)

Handbook	N	ICC	95% CI Lower Bound	95% CI Upper Bound
Elementary Literacy	357	0.65	0.59	0.71
Elementary Mathematics	363	0.69	0.64	0.74
Secondary English-Language Arts	266	0.61	0.52	0.68
Secondary History/Social Studies	253	0.75	0.69	0.80
Secondary Mathematics	190	0.70	0.62	0.77
Secondary Science	214	0.61	0.52	0.69
Physical Education	133	0.65	0.55	0.74
Special Education	489	0.63	0.57	0.68
Early Childhood	232	0.75	0.69	0.80
Visual Arts	66	0.69	0.54	0.80
Middle Childhood Mathematics	53	0.46	0.22	0.65
Middle Childhood History/Social Studies	36	0.78	0.60	0.88
K–12 Performing Arts	153	0.65	0.55	0.73
Elementary Education: Literacy with Mathematics Task 4	804	0.68	0.64	0.72
Elementary Education: Mathematics with Literacy Task 4	36	0.70	0.49	0.84

Internal Structure

The use of edTPA rubrics, tasks, or overall scores depends on the intended purpose as well as the policy and approach to implementation of each program and state. The score on a particular rubric provides a candidate's level of readiness on the particular skill/ability being measured and informs conversations about the strengths and opportunities for growth of a particular candidate or a preparation program. Scores on each rubric and total scores for the three edTPA tasks are reported to candidates, programs, and states to inform decisions and level of competency for each of the three components of the teaching cycle (i.e., planning, instruction, and assessment). The final score is the summed score across rubrics in all three tasks and is used as an overall measure of readiness to teach. As a licensure assessment, the claim is made that the scoring procedure appropriately summarizes relevant aspects of performance and is applied accurately and consistently for all candidates. This is based on evidence that the scoring rules are appropriate and that the data fit the scoring model. The analyses of the internal structure of edTPA provide psychometric evidence that supports the structure of levels within each rubric, the fit of rubrics within the three edTPA tasks, and the use of a single summed total score to represent candidates' overall performance. The accuracy and consistency of the scoring process is supported by the scoring model, scorer training, double scoring procedures, and quality management outlined in previously published administrative reports.

Confirmatory Factor Analyses

Exploratory factor analyses of 2013 field test data provided support for the use of a total score on edTPA to summarize a candidate's performance, as well as for the underlying task structure (see pg. 22 of the edTPA 2013 Field Summary Report). Confirmatory Factory Analyses (CFA) test whether patterns (correlations) among observed scores on a set of test items conform to hypothesized structures (Brown, 2006), providing validity evidence based on a test's "internal structure" to support score interpretations (AERA, APA, & NCME, 2014).

These analyses included all 2019 first-time edTPA submissions and excluded incomplete portfolios and portfolios with condition codes.[1] In cases where a portfolio was double scored, only the first rater's score was included in the analyses. CFA models were estimated based on the observed sample covariance matrix among rubric scores for the 2019 administration cycle. All models were estimated via maximum likelihood as implemented in the R package "lavaan" (Rosseel, 2012), which also generates standard errors and scaled chi-square fit statistics. Based on the design and interpretation of the edTPA **total score**, a 1-factor model in which all rubric scores load on a single latent factor was estimated. To account for the task-based design and structure of edTPA portfolios, a 3-factor model with correlated factors and with each rubric loading only on its associated task (i.e., rubrics 1–5 were assigned to "planning"; rubrics 6–10 were assigned to "instruction"; rubrics 11–15, were assigned to "assessment") was also estimated. All factor loadings for the three-factor solution (i.e., "planning," "instruction," and "assessment") were positive and statistically significant as anticipated (most standardized loadings were greater than .6 in the 3-factor model)—meaning that there was reasonable agreement in the data regarding what the rubrics were designed or intended to measure and their relationship to that construct. Table A in Appendix A presents the estimated standardized factor loadings for the 1- and 3-factor models in the full sample of portfolios. Table B presents the estimated correlations among the task factors in the 3-factor model, which are also strongly positive and statistically significant. The large magnitude of the correlations further supports the interpretation that edTPA rubrics

[1] Condition codes are applied to one or more rubrics when the candidate's materials do not comply with edTPA evidence requirements (e.g., inaudible video, missing artifact, wrong artifact) and are therefore unscorable.

measure three highly interrelated sub-dimensions—planning, instruction, and assessment—of a single readiness-to-teach construct.

IRT: Partial Credit Model

As outlined in prior administrative reports, a polytomous item response theory (IRT) model, the partial credit model (PCM; Masters, 1982), was fit to the same sample of edTPA submissions included in the CFA models. The PCM provides a statistical model of the probability that a candidate earns each possible rubric score as a function of a single, continuous, underlying dimension "theta."

The unidimensional PCM was fit to the 2019 sample of candidates. Models were estimated using marginal maximum likelihood. As noted above, edTPA scores are derived from the ratings of a single scorer who scores the entire portfolio (i.e., multiple scoring rubrics) of each candidate. Additionally, the results presented below are based on aggregating data across credential areas. Because edTPA is based on a single total score calculated equally across fields with 15 rubrics, this analysis provides evidence about how this total measure functions overall.

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. Plots of expected and observed rubric scores across the theta range were compared across the theta range to identify potential model misfit. A Wright Map depicting the distribution of candidate proficiency estimates alongside rubric threshold parameter estimates was inspected to determine (a) whether rubric thresholds conformed to the expected ordering and (b) whether the rubric thresholds for each score point were well-distributed across the range of the theta distribution. Finally, to summarize precision of theta estimates, the test information function and conditional standard error of estimate were plotted across the range of the theta distribution, and a person separation reliability index was estimated.

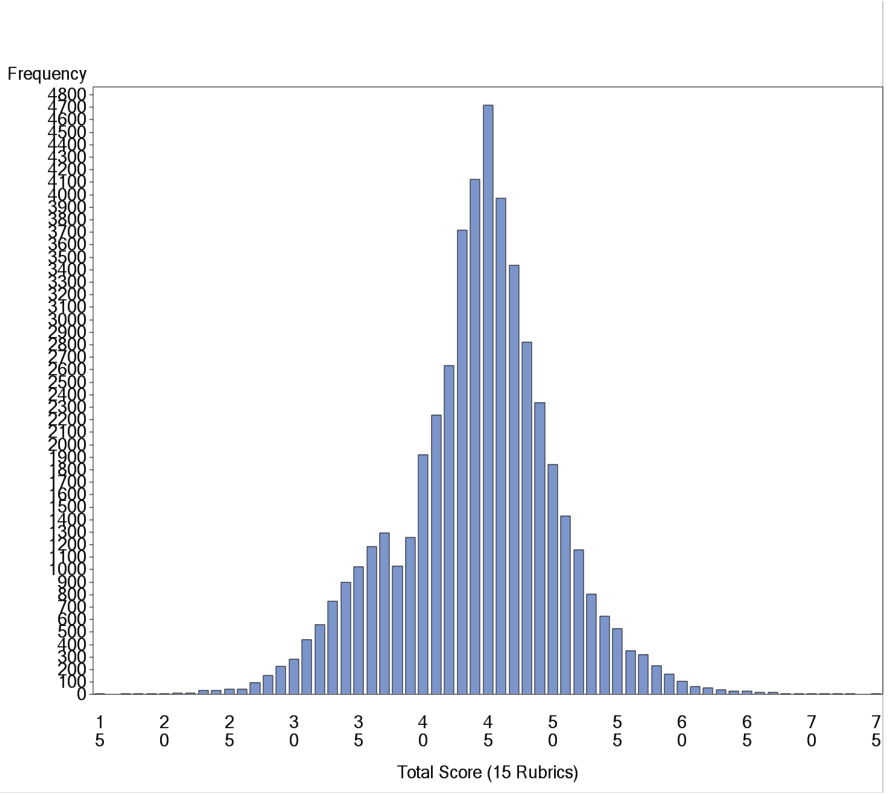
Inspection of the Wright Maps and rubric parameter estimates showed the hypothesized ordering of rubric thresholds and demonstrated that the Thurstonian thresholds (proficiency level at which a candidate has a 50% chance of scoring above a given score level) were located across the entire range of estimated candidate performance on the theta scale (see Appendix A, Table A-3). The test information function (and hence standard error of measurement in the theta metric) was consistent across the range of candidate performance.

To summarize, these results provide information about the level of performance at which candidates are likely to move from one possible rubric score to the next. The fact that these points are distributed across the theta distribution affirms that edTPA rubrics are constructed to provide useful discriminating information about candidate performance at different levels of overall performance.

Candidate Performance

Overall Scores

The following figure presents the score distribution of 49,123 edTPA portfolios, in fields scored based on 15 rubrics and submitted January 1 to December 31, 2019, the sixth full calendar year for which edTPA was used consequentially.



This represents the distribution of final scores on all complete portfolios scored on five separate rubrics within each of the three major edTPA tasks: planning, instruction, and assessment. There are five levels of possible performance for each rubric, with level 3 characterizing “ready to teach” and a total score range from 15 to 75. This figure shows that scores are approximately normally distributed across this range. Figures presenting further information on the distribution of these portfolios (distribution based on first score only, and distribution within cut band) are found in Appendix B.

Task and Rubric Scores

Summary descriptive statistics and distributions for each task and rubric for the 49,123 edTPA portfolios, in fields scored based on 15 rubrics and submitted January 1 to December 31, 2019, are presented in the following table, including mean, standard deviation, standard error of the mean, and minimum/maximum score obtained.* As a reference, rubrics are listed below by title.[2]

RUBRIC	MEAN	S.D.	SE OF MEAN	MIN	MAX
Task 1: Planning					
P01	3.0	0.6	.00	1	5
P02	3.0	0.7	.00	1	5
P03	3.1	0.6	.00	1	5
P04	2.9	0.6	.00	1	5
P05	2.9	0.7	.00	1	5
Task 1 Total	**14.9**	**2.4**	**.01**	**5**	**25**
Task 2: Instruction					
I06	3.1	0.4	.00	1	5
I07	3.0	0.6	.00	1	5
I08	2.9	0.6	.00	1	5
I09	2.9	0.7	.00	1	5
I10	2.7	0.6	.00	1	5
Task 2 Total	**14.5**	**2.0**	**.01**	**5**	**25**
Task 3: Assessment					
A11	2.9	0.8	.00	1	5
A12	3.2	0.8	.00	1	5
A13	2.7	0.7	.00	1	5
A14	2.9	0.7	.00	1	5
A15	2.9	0.7	.00	1	5
Task 3 Total	**14.7**	**2.8**	**.01**	**5**	**25**
Overall Total	**44.2**	**6.1**	**.03**	**15**	**75**

Task 1: Planning

P01. Planning for Content Understandings
P02. Planning to Support Varied Student Learning Needs
P03. Using Knowledge of Students to Inform Teaching and Learning
P04. Identifying and Supporting Language Demands
P05. Planning Assessments to Monitor and Support Student Learning

Task 2: Instruction

I06. Learning Environment
I07. Engaging Students in Learning
I08. Deepening Student Learning
I09. Subject-Specific Pedagogy
I10. Analyzing Teaching Effectiveness

Task 3: Assessment

A11. Analysis of Student Learning
A12. Providing Feedback to Guide Learning
A13. Student Understanding and Use of Feedback
A14. Analyzing Students' Language Use and Content Learning
A15. Using Assessment to Inform Instruction

[2] Descriptive statistics for Task 4 rubrics of the Elementary Education: Literacy with Mathematics Task 4 (M19, M20, M21) and Elementary Education: Mathematics with Literacy handbooks (L19, L20, L21) are presented in Appendix C.

*Does not include 13-rubric fields (World Language, Classical Languages); includes first 15 rubrics of Elementary Education only.

The average edTPA score across 49,123 portfolios from fields with 15-rubric handbooks was 44.2, with a standard deviation of 6.1. This average performance shows only a slight decrease (from 44.3 in 2018) in mean scores and reduction in the standard deviation. As a reference, the 2013 field test data average score was 42.8 (SD = 8.17). As per previous years, scores ranged across the entire range of possible scores, from 15 to 75.

Candidate performance by task shows that candidates performed most highly on the planning task (mean = 14.9) followed by the assessment tasks (mean = 14.7) and scored the lowest on the instruction tasks (mean = 14.5).

Performance by Content Field

Distribution of total scores for all fields (and sample size for each field) can be found in Appendix C.

In 2013, SCALE, in collaboration with a standard-setting panel consisting of stakeholders from both K–12 and higher education communities established a Recommended Professional Performance Standard of 42. To date no state has established a cut score higher than 41. Trend data show mean performance in the content areas of Health Education, Physical Education, and Secondary Mathematics Education mean performance are lower than 41—the highest performance standard in use during 2019. An examination of data over the past four years indicates a downward trend of performance for both Health Education and Physical Education, while the data for Secondary Mathematics is slightly varied over the same time frame. SCALE and Evaluation Systems continue to examine the data to identify trends and gather input from the field to ensure alignment between the handbooks and the content area.

Due to differences in sample size, content knowledge demands, and low numbers of submissions in some fields, direct comparisons across fields should be approached with caution.

All edTPA handbooks examine the same underlying constructs and follow the same architecture. An analysis by SCALE of edTPA rubrics revealed an 85.7% overlap of common language across rubrics, across content areas. Although several rubrics have subject-specific language embedded within the rubric criteria, the construct of the rubric remains identical. Rubrics with the same construct but only varying by subject-specific phrases embedded within the level descriptors are noted in the following table. Most rubrics had an 80% overlap in language with the exception of rubric 9 – Subject-Specific Pedagogy. Rubric 9 is a subject-specific rubric designed to assess subject-specific constructs identified by the design team for each content area. Patterns of performance across content fields are confirmed systematically in a multipronged approach:

1. Factor analyses models of latent structure are reviewed for each field.
2. Summary data, distributions, and patterns of overall scores, tasks, and rubrics are compared across fields for flags of outlier behavior.
3. Indices of reliability (internal consistency, exact and adjacent agreement by rubric, kappa-Ns) are reviewed for each field
4. Scoring trainers and supervisors are consulted to confirm scoring and backreading processes and flag any recurring questions from scorers, including recalibrating or intervening as needed.
5. Experts in each profession are consulted to review the data, and to discuss alignment of the handbook to the standards and expectations of the profession.
6. Input from programs and faculty using edTPA via the Online Community at edTPA.AACTE.org and emails to SCALE or AACTE staff are reviewed.
7. Review and clarification of handbooks, scorer training, and support materials is conducted annually based on all quantitative and qualitative data.

Performance by Consequential Use

edTPA portfolios officially scored in the 2019 operational year represent submissions from candidates in 32 states. Of the 49,123 portfolios scored (excludes 13-rubric handbooks, 10 submissions from the Educational Technology Specialist and Literacy Specialist handbooks due to low Ns, and includes the first 15 rubrics of the 18-rubric handbooks), 7,311 were submitted by candidates in states that do not currently have policy for edTPA use, and 41,822 were submitted in states with consequential policy. States without policy with submissions in 2019 are AZ, CO, IN, KY, NE, OH, OK, PA, TX, UT, VA, and WY. States with consequential policy and submissions in 2019 are AL, AR, CA, CT, DE, GA, HI, IL, IA, MD, MN, NC, NJ, NY, OR, SC, TN, WA, WV, and WI. Appendix D presents the approximate percentage of portfolios coming from each state, and Appendix E shows overall performance (mean, standard deviation, and number of submissions) by field in states without state-wide policy for use of edTPA, and states where such policy exists.

State policy mandating edTPA for purposes of teacher licensure and/or program evaluation results in greater consistency for use and implementation. It was therefore hypothesized that submissions from states with official policy would have higher average scores than those from states without edTPA policy. As predicted, edTPA scores were higher in states with policy requiring edTPA, with policy states having a mean of 44.5 and non-policy states having a mean of 42.3. This finding is consistent with expectations given the increased consistency of implementation and support structures, as well as levels of effort and motivation, that come about as a result of state-wide policy for consequential assessment.

Modeled Candidate Passing Rates

The following table reports the percentage of candidates (out of 49,123) who would have "passed" edTPA (based on edTPA 2019 data) at different potential cut scores for edTPA assessments with 15 rubrics. The table lists possible passing scores within the band of 35 and 42 for candidates in states with and without consequential policy. Estimated passing rates are reported for cut scores within this band. These passing rates are pooled across states and credential areas. **Note that these data include portfolios submitted in states where edTPA was not completed under high-stakes or consequential circumstances, and from institutions that may still be in the process of developing support strategies for candidates.** Passing rates by program and state are likely to differ based on policy, support structures, and experience with edTPA.

Cut Score	Modeled Overall Passing Rate
35	92.7%
36	90.6%
37	88.2%
38	85.5%
39	83.5%
40	80.9%
41	77.0%
42	72.4%

Note that to date, passing scores established by states range from 35 to 41 (for 15-rubric handbooks).

See the State edTPA Policy Overview for the most up-to-date information on state policies and any established state consequential score information, including planned adjustments over time.

Performance by Demographic Groups

When submitting an edTPA portfolio for official scoring, the candidate is asked to provide self-reported demographic information in several categories: gender, ethnicity, teaching placement context, program type, education level, and primary language. Candidates are provided with the option to "Decline to Answer." **Analyses of performance by groups within these categories included only 15-rubric handbook portfolios submitted in states that have policy for consequential use of edTPA (total N = 41,822).**

The analyses revealed small differences in performance across some of the groups. **It is important to note that the difference in sample sizes of some of the groups within each demographic category may affect the ability to generalize these results to the teacher candidate population;**

all estimates of performance should not be overgeneralized and should be interpreted with caution. Further, differences in performance do not take into account any prior or experiential differences in the applicant pool, differences in program quality or preparation of candidates, K–12 differences in candidate support, and other factors that may contribute to a candidate's score and cause differences in performance. Lastly, a difference in mean score may not necessarily influence pass rates at various cut scores required by individual states. See the section below for modeled pass rates by demographic group by cut score.

What follows is a description of the performance in the following categories: Teaching Context (School Locale and Teaching Placement), Ethnicity, Primary Language, Gender, and Education Level. Finally, a regression analysis was conducted to examine the contribution of these demographic categories in explaining and interpreting edTPA candidate scores.

Teaching Context (School Locale)

Upon submission of their edTPA portfolio, candidates are asked to indicate the context of their teaching placement. Based on these data, an ANOVA was run to analyze whether overall edTPA scores differed based on the teaching context of the candidate. The table below displays mean scores, standard deviations, and submission volumes by teaching placement categories. Results showed that candidates teaching in suburban and city settings had the highest average scores, while candidates teaching in town settings had the lowest average scores. For the ANOVA and Games-Howell post hoc analyses, see Appendix F.

School Locale	N	Mean	Std Dev
City	15,653	44.61	6.01
Rural	6,815	44.37	5.88
Suburb	12,967	44.73	5.82
Town	5,911	43.93	6.04
Declined to answer	476	42.95	6.45

This finding provides evidence that candidates in suburban settings had the highest overall performance and that candidates whose practice or clinical teaching takes place in town settings have lower average scores. While the difference in performance did reach significance ($p < .05$), the practical difference is less than one score point. For mean difference analyses see Appendix F.

Teaching Context (Student Characteristics)

The current report also presents a set of teaching context variables associated with student characteristics in the candidate's placement, including the percentage of English language learners, students receiving free and reduced-price lunch, and students eligible for special education services, as shown in the tables below.

% ELL Students	N	Mean	Std Dev
<10%	24,134	44.51	5.89
10–25%	6,524	44.63	5.88
26–50%	3,119	44.67	5.89
51–75%	1,389	44.18	5.93
76–100%	2,023	44.75	6.74
Declined to answer	4,633	44.04	5.97

% Students Free and Reduced-Price Lunch	N	Mean	Std Dev
<10%	6,159	44.40	5.85
10–25%	5,348	44.80	5.86
26–50%	5,834	44.82	5.88
51–75%	5,349	44.53	6.02
76–100%	11,421	44.43	6.02
Declined to answer	7,711	44.17	5.94

% Students SPED Eligibility	N	Mean	Std Dev
<10%	17,778	44.61	5.81
10–25%	11,065	44.47	6.01
26–50%	3,276	44.47	5.94
51–75%	629	43.86	5.94
76–100%	3,475	44.57	6.31
Declined to answer	5,599	44.17	5.99

When considering the number of ELL students in the classroom, candidates reporting 76–100% ELL students performed the highest (mean = 44.75). Candidates reporting 51–75% ELL students in the classroom had the lowest performance (mean = 44.18). The practical mean difference between the highest and lowest mean performance is .57 score points; this difference did not reach significance ($p > .05$).

Considering the number of students eligible for free and reduced-price lunch (FRPL) the highest performance (mean = 44.82) is by candidates reporting 26–50% FRPL and the lowest performance (mean = 44.40) by candidates reporting < 10% FRPL. The score differences in the highest and lowest performance for both ELL and FRPL are less than one score point (0.42). This group difference did not reach significance ($p > .05$).

Performance for candidates reporting fewer students who are SPED eligible (< 10%) outperform candidates who report higher numbers of students who are SPED eligible (51–75%) with mean performances of 44.61 and 43.86, respectively, and result in a practical score difference of less than one score point (significant at $p < .05$).

Overall, performance related to teaching contexts (student characteristics) remained stable. Mean performance scores from 2018 are similar to the current findings, with point differences varying by less than one score point.

Although the differences in candidate performance for these variables are very small (in most cases less than one score point), the analysis of student characteristics variables revealed some statistically significant differences. In general, candidates teaching in rural, suburban, and city placements outperformed candidates teaching in town placements; variable differences are found in performance related to the percentage of ELL students, the percentage of students eligible for free and reduced-price lunch, and the percentage of students who are eligible for SPED services.

Teaching Context (Placement Length)

edTPA candidates are asked to indicate the amount of time spent in a clinical/student teaching placement prior to completing edTPA. As can be seen below, there were differences in candidate performance associated with their reported length of placement completion (shown in number of weeks) prior to edTPA. Strongest candidate performance was found in placement completion of 21–30 weeks, and lowest performance was indicated in placements of zero weeks. While this was significant ($p < .05$) the score difference was 2.28 score points. Unequal N for these two groups restrict generalizations from these data. Note: Candidates assigned as the teacher of record may have selected "0" in response to this question.

Length of Placement (weeks)	N	Mean	Std Dev
0	409	42.69	6.72
1–10	9,701	44.49	6.00
11–20	10,381	44.50	5.77
21–30	5,822	44.97	5.88
31–40	3,306	44.71	5.91
41+	7,947	44.38	6.03
Declined to answer	4,256	44.01	6.06

Different teaching contexts present different sets of experiences and opportunities for teacher candidates. Many programs purposefully place students in field experiences in a range of teaching contexts and vary in their approach to preparing candidates for teaching in different contexts. Therefore, depth and breadth of experiences provided by the preparation program should be considered. The process used by each program to select

candidates, the resources and supports available, as well as the candidate's disposition or preference are also likely to play a role in how a candidate performs within a particular teaching context. These data can help programs to reflect on how they serve and prepare their candidates and to scaffold conversation about teaching strategies that best support learners across various teaching contexts.

Program Type

Candidates registering for edTPA were asked to identify which type of program best described their preparation. The table below shows the means for various options that distinguish traditional university-based programs from alternative certification in a university, district, or Teach for America.

Teacher Preparation Route	N	Mean	Std Dev
Traditional EPP	35,320	44.64	5.83
Teach for America	693	44.62	6.92
Alt—university or college-based program	3,567	43.74	6.42
Alt—district provider	899	43.33	6.58
Other	562	42.55	6.88
Declined to answer	781	43.70	5.64

Strongest performance was seen by candidates who indicate they were enrolled in a traditional educator preparation program. Of candidates who selected an identified teacher preparation route, lowest performance was seen by candidates who indicated preparation by an Alt—district provider. While this did reach significance ($p < .05$), the score difference is 1.31 points. Unequal N for each program option and small numbers in alternative certification categories restrict generalizations from these data.

Ethnicity

Data from the 2019 operational year indicated that the large majority of candidates submitting edTPA portfolios were White (73.9%), followed by candidates identifying as Hispanic (8.4%), African American/Black (6.5%), Asian or Pacific Islander (4.2%), and American Indian or Alaskan Native (< 1%), with 3.3% identifying as Multiracial, 1.2% Other, and 2.1% not identifying ethnicity. **The disproportionate representation of White candidates and the relatively small sample sizes of all other candidate groups must be interpreted with caution when making comparisons or generalizations to sample sizes or to the general population of teacher candidates.**

The table below shows the sample size, percentage of sample, average scores, and standard deviations of each group.

Ethnicity	N	% of total N	Mean	Std Dev
African American/Black	2,726	6.5%	42.74	6.11
American Indian or Alaskan Native	141	<1%	42.99	6.51
Asian or Pacific Islander	1,744	4.2%	44.63	5.94
Hispanic	3,508	8.4%	44.42	5.76
White	30,910	73.9%	44.69	5.89
Multiracial	1,390	3.3%	44.35	6.08
Other	517	1.2%	44.12	6.18
Undeclared	886	2.1%	43.67	6.55

Analyses revealed that average scores for Asian or Pacific Islander, Hispanic, White, and Multiracial candidates were quite similar, and observations over time from past administration years have shown Hispanic candidate performance (the second largest ethnicity group) and Multiracial candidates closely mirroring that of White candidates. While the average scores of African American/Black and American Indian or Alaskan Native candidates were lower than those of other candidate groups ($p < .05$), the fact that African American/Black candidates made up a very small portion of the candidate pool (6.5%) and the low N for American Indian or Alaskan Native candidates (141) should be considered and interpreted with caution.

To determine whether the scores of two groups are meaningfully different from one another, it is informative to compare the difference in means of the

two groups to their pooled standard deviation. A smaller ratio indicates that there is substantial overlap in the scores of the two groups, and greater variability within each group than between them. The difference in means between White and African American/Black candidates is 1.95 points, and the pooled standard deviation is the square root of $(5.89^2+6.11^2)/2 = 6.00$. The difference in the mean performance of African American/Black and White candidates in this sample, then, is about one-third (0.325) of a standard deviation (1.95/6.00). Similarly, the difference between White and American Indian or Alaskan Native candidates is 1.70, and the pooled standard deviation is the square root of $(5.89^2+6.51^2)/2 = 6.21$. The difference in the mean performance of American Indian or Alaskan Native and White candidates in this sample, then, is just over one-quarter (0.274) of a standard deviation (1.70/6.21).

These findings contextualize the magnitude of the difference; placing these findings in the context of assessment of teacher education, the gap between average scores of White candidates and other candidate groups is smaller than that seen with other more traditional standardized assessments of initial teaching (e.g., Goldhaber and Hansen, 2010).

Trend data in the performance between White candidates and other candidate groups in the 2016 and 2019 data suggest a slight narrowing of a gap in scores. The difference in mean score for White candidates compared to African American/Black candidates for the 2018 and 2019 reporting periods has decreased by 0.26 score points. Over the period of 2016–2019, a gap in scores between White and African American/Black candidates has decreased by more than half a score point (0.57).

Performance between White and American Indian/Alaskan Native candidates decreased by 0.71 over the 2018 and 2019 periods. When considering the differences in performance between White and American Indian/Alaskan Native candidates, in 2016 White candidates outperformed American Indian/Alaskan Native candidates by 3.75 points; in 2019 that difference was 1.70—a change in score differences between these groups of 2.05 score points. Taken together, these may signal a narrowing of a gap. We are encouraged by these findings and continue to explore the impact of ethnicity on candidate performance.

The performance of candidates was also examined by ethnic group within each teaching context to see whether the pattern of the overall sample was consistent across the different placements (contexts). This examination revealed that difference in means of the White and Hispanic candidates and White and Multiracial candidates was less than 1 point and varied across teaching contexts. **Furthermore the examination also revealed a gap between mean scores for Hispanic candidates and White candidates has decreased from 0.43 (2018) to 0.27 in 2019 and no longer represents a statistically significant difference ($p > .05$). Similarly, the difference in mean scores between Asian candidates and Hispanic candidates is no longer statistically significant ($p > .05$). We are encourage by this narrowing of a gap for these candidate groups.** In a large data set with varied differences in sample size and mean scores among various demographic variables (e.g., ethnicity, gender) significance tests tell only one part of the story. In addition to examining mean differences in the overall edTPA score, states and programs should examine the impact these small overall score differences may or may not have on pass rates using their required cut score. The table below reveals the percentage pass rate by group by cut score. To date, only CA, IA, IL, TN, and WA have a passing standard above this score.

These data reveal overall trends in edTPA performance for this sample, and while findings should not be overgeneralized, educator preparation programs and state agencies are encouraged to use data from their respective populations to conduct further analyses and consider implications. edTPA is committed to providing an equitable assessment that is free of bias and adverse impact.

Pass Rates (in %) by Demographic Communities												
		Gender			Ethnicity							
Cut Score	All	Not Provided	Male	Female	African American/ Black	American Indian or Alaskan Native	Asian or Pacific Islander	Hispanic	White	Multiracial	Other	Undeclared
35	93	92	89	94	87	90	94	94	93	93	92	89
36	91	91	86	92	85	88	92	91	91	91	90	87
37	88	90	83	90	81	84	91	89	89	88	86	84
38	86	85	80	87	78	80	88	87	86	85	83	81
39	83	84	77	85	75	78	86	85	84	83	81	79
40	81	82	74	83	72	77	83	82	82	80	78	77
41	77	76	70	79	67	72	79	79	78	75	74	72
42	72	70	65	74	61	67	73	74	73	71	70	67

Note: Based on 2019 examinees in all states with a total National score based on 15 rubrics

Primary Language

Candidates were asked to identify whether English is their primary language. Primary language English candidates had a mean score higher than those candidates who indicated their primary language was not English ($p < .05$), with a score difference of less than one score point (.62). Note: The disproportionate N size for these two populations would warrant caution in overgeneralizing (candidates who indicated their primary language was not English represent only 2.73% of the population).

Primary Language	N	% of N	Mean	Std Dev
English	40,269	96.28%	44.51	5.94
Other	1,141	2.73%	43.89	5.96
Declined to answer	412	.98%	43.97	6.02

Gender

In this sample, 77.5% of submissions came from female candidates and 21.72% from male candidates. Female candidates scored higher than their male counterparts; this difference (1.55) was statistically significant ($p < .05$).

Gender	N	% of N	Mean	Std Dev
Not Provided	308	.74%	44.24	5.85
Male	9,083	21.72%	43.28	6.27
Female	32,431	77.55%	44.83	5.81

Education Level

The achieved level of education completed prior to taking edTPA was reported by candidates. With the exception of the Master's/Master's plus credits level, mean values are the same to one decimal point. Mean comparison between Master's/Master's plus credits and both High school/Some college and Bachelor's/Bachelor's plus credits reached significance ($p < .05$).

Due to the different group sizes, results should be interpreted with caution; this is particularly true for the Masters/Master's plus credits and Doctorate levels. Structure of program curricula, timing of the assessment within the program, and prior experience with pedagogical theory and teaching practice may also play a role in outcomes on an assessment of teaching readiness.

Education Level	N	Mean	Std Dev
High school/Some college	19,548	44.53	5.74
Bachelor's/Bachelor's plus credits	18,131	44.55	6.07
Master's/Master's plus credits	3,943	44.00	6.34
Doctorate	200	44.56	6.59

Regression Analysis

Regression analyses are used to determine whether particular variables significantly predict an outcome, and the extent to which these variables explain the differences in outcomes within the sample. To examine the contribution of all demographic factors to the performance of the candidates, a multiple regression model including school context, program type, student characteristics, ethnicity, gender, education level, and primary language was run to examine the extent to which demographic factors explain the variability in total edTPA scores.

It is important to note that finding that a factor is a "statistically significant" predictor does not necessarily indicate that this factor makes a substantial or meaningful difference in outcomes. This is particularly true given the large number of candidates used in the analysis. The percent of variance explained (R^2) by each factor is therefore presented here to describe the extent to which each variable explains the differences in candidates' edTPA scores.

The overall model was statistically significant—F(47,41774) = 27.21 ($p < 0.001$)—indicating that this model predicts edTPA scores better than chance alone. The following table presents each factor included in the model and the percentage of variance in total scores accounted for by each factor and by the overall model.

Factor	Variance Explained—R^2(%)
Teacher Preparation Route	0.45
School Locale	0.26
Gender	1.11
Ethnicity	0.65
Primary Language	0.03
Education Level	0.19
Length of Placement (weeks)	0.12
Free and Reduced Lunch	0.08
Special Education Eligibility	0.04
English Language Learners	0.04
Overall Model	2.971

Note: The percent of variance explained is actually the semi-partial eta-square statistic that shows the unique contribution of the factor to total variance (SAS 2017, pg. 3786).

Some notable changes in the variance explained by specific factors from 2018 to 2019 include School Locale, Gender, Ethnicity, Length of Placement, and Special Education Eligibility. The largest change in variance in scores explained by the factors is attributable to Gender, which has increased this year, with 1.11% of the variance in scores, as compared to .66% in 2018. While this is notable, Gender still represents only 1.11% of the reasons candidates scored as they did. We are encouraged that in the 2019 model, changes in variance attributable to School Locale, Ethnicity, Education Level, Length of Placement, and Special Education Eligability were all reduced as compared to the 2018 model.

This model accounts for only 2.97% of the variance in total scores (R2 = .0297); 97.03% of the variability in scores is explained by other factors not accounted for by the variables included in this model. Almost all statistical tests are significant when the sample sizes are large. Hence the real meaning of a difference must rely on the judgment in the user regarding any "practical significance" of such differences. The differences seen in this table are very small in terms of percentage of variance explained and may not be practically meaningful. In other words, a candidate's demographic characteristics alone are a poor predictor of a candidate's edTPA performance or readiness to teach despite statistical significance. This finding further supports the conclusion that while some statistically significant differences do exist between candidate groups, more than 97% of the explanation of candidate performance can be explained by other non-demographic factors. How a candidate performs on edTPA may be largely explained by other factors such as the candidate's knowledge, skills, and abilities to begin teaching; initial level of academic readiness; the quality of the preparation program; and/or the supports provided by the program. Further research into each of these and other variables can serve to inform the ways in which candidates, faculty, and programs employ edTPA as a tool for candidate and program learning.

Conclusion

As with the field test data (2013) and those of the previous administrative reports data presented here are intended to meet the technical standards of APA, AERA, and NCME (2014) and support the use of edTPA to grant an initial license to teacher candidates as well as to inform state and national accreditation. The reporting of performance of all candidates who submitted edTPA portfolios in 2019 is presented for all content fields and informs the use of edTPA across states.

As is the case with NBPTS, educative use of a performance-based assessment is more than a testing exercise completed by a candidate. edTPA's emphasis on support for implementation mirrors the NBPTS's use of professional networks of experienced users to assist others as they prepare for the assessment. The opportunities for educator preparation program faculty and their P–12 partners to engage with edTPA are instrumental to its power as an educative tool. The extensive and growing library of resources developed by

SCALE, the National Academy of consultants, and state infrastructures of learning communities for faculty and program leaders promote edTPA as a tool for candidate and program learning. As candidates are provided with support opportunities to develop and practice the constructs embedded in edTPA throughout their programs, and reflect on their edTPA experience with faculty and P–12 partners, they are more likely to internalize the cycle of effective teaching (planning, instruction, and assessment) as a way of thinking about practice—a way of thinking about students and student learning that will sustain them in the profession well beyond their early years in the classroom.

Appendix A: Internal Structure

The table below presents the estimated standardized factor loadings for the 1- and 3-factor models in the full sample of consequential state portfolios (N = 41,822) for all 15-rubric handbooks, including the first 15 rubrics of the 18-rubric handbooks.

Table A-1: Confirmatory Factor Analyses: Standardized Factor Loading Estimates

	1-Factor Model	3-Factor (Task) Model		
Rubric	**F1**	**Planning**	**Instruction**	**Assessment**
1	0.596	0.652	--	--
2	0.570	0.636	--	--
3	0.599	0.637	--	--
4	0.559	0.594	--	--
5	0.588	0.660	--	--
6	0.430	--	0.516	--
7	0.592	--	0.713	--
8	0.561	--	0.685	--
9	0.564	--	0.648	--
10	0.573	--	0.555	--
11	0.641	--	--	0.681
12	0.528	--	--	0.597
13	0.592	--	--	0.673
14	0.604	--	--	0.624
15	0.635	--	--	0.691

Note: "--" indicates factor loadings fixed to 0.0 in model estimation.

All factor loadings for the 3-factor solution were positive, and most were statistically significant as anticipated. Only four of the factor loadings in the 3-factor model were less than 0.60.

Table A-2: Confirmatory Factor Analyses: Task Factor Correlation Matrix

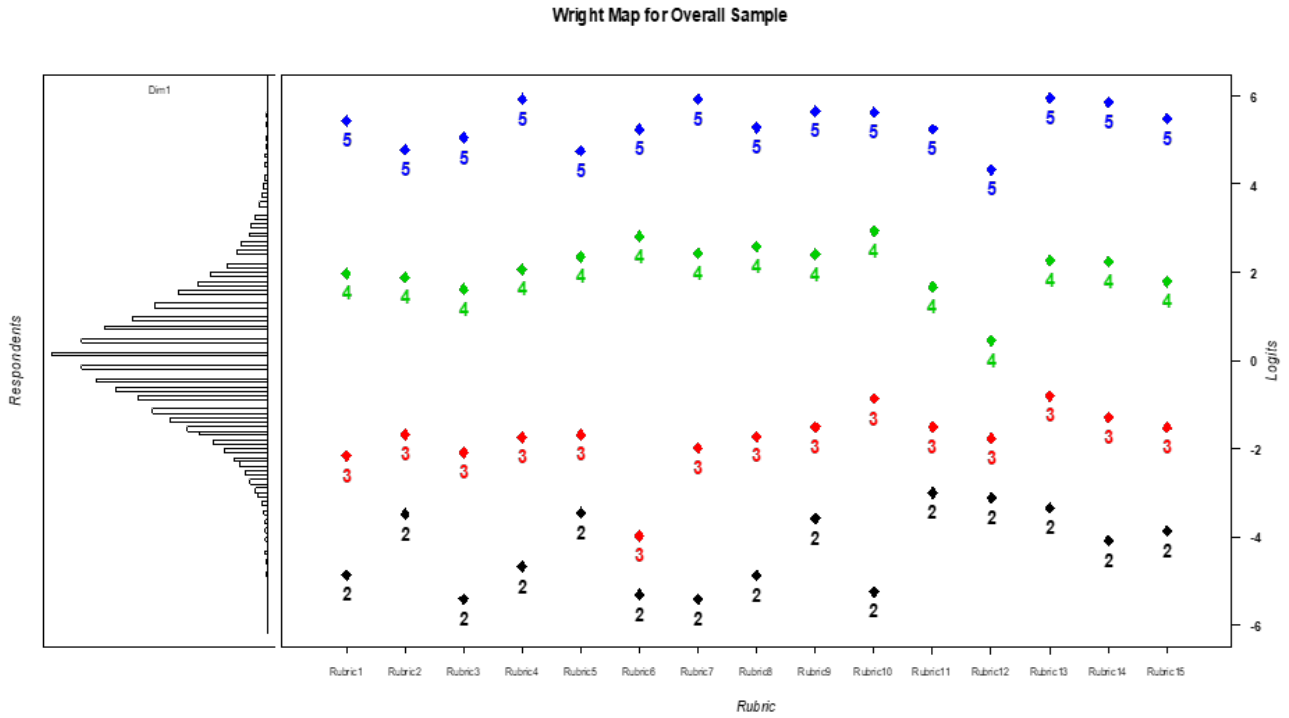
The table below presents the estimated correlations among the task factors in the 3-factor model.

	Planning	Instruction	Assessment
Planning	1.000	--	--
Instruction	0.724	1.000	--
Assessment	0.775	0.704	1.000

The task factor correlations are moderately strong and statistically significant. This result supports the edTPA structure consisting of three correlated abilities: Planning, Instruction, and Assessment. Since each rater evaluated all 15 rubrics for a portfolio, it is uncertain how much this contributes to the correlated factors displayed.

Table A-3: Partial Credit Model: Wright Map

The following figure shows the ordering and distribution of Thurstonian thresholds across the range of candidates' theta estimates. The histogram on the left shows the distribution of candidate theta estimates (analogous with total performance). These are a direct function of total scores, which represent estimates of teacher readiness. The points on the graph (Thurstonian thresholds) represent the point on the underlying scale at which a candidate has a 50% chance of scoring at or above score k for a particular rubric. For example, the furthest left point labeled "2" indicates the point on the scale at which a candidate is predicted to have a 50% chance of scoring a 2 or higher on Rubric 1, the furthest left point labeled "3" is the point at which a candidate is predicted to have a 50% chance of scoring a 3 or higher on Rubric 1, and so on.



This graph shows that the ordering of thresholds is as intended (the threshold for scoring 3 is higher than for scoring 2 on a given rubric, etc.). This graph also shows that thresholds are typically evenly distributed across the theta distribution, indicating that differences in rubric scores are sensitive to differences in candidate performance at a range of performance levels. Rubric 6 shows a lower threshold for scoring 3 than other rubrics.

Appendix B: Performance by Content Field

The following tables contain average candidate performance—overall and for each task and rubric—for 15-, 13-, and 18-rubric content fields.

Data cautions for interpretation of tables:

- Total portfolio scores are based on the 13, 15, or 18 rubrics (depending on the handbook) that are common to all national handbooks.
- The difference in the number of Elementary Education Handbook portfolios is a result of the exclusion of examinees who received condition codes for Task 4 from the 18-rubric handbooks data summary (see Table B-3: Average candidate performance for 18-rubric content fields); these examinees were included in the data summary in Table B-1: Average candidate performance for 15-rubric content fields within either the Elementary Literacy or Elementary Mathematics fields, depending on the Elementary Education handbook that was used.
- Results for Washington handbooks are included in the national results reported here and are based on the rubrics common to all handbooks. State-specific rubrics, such as Washington's Student Voice, are excluded for the purpose of this report.
- Occasionally, rubrics receive a final score ending in a ".5". This occurs when edTPA portfolio submissions are scored by two independent scorers. For those portfolios, the final rubric score is the average of the scores assigned by each scorer.
- For this report, the scores included in the distribution of portfolio total scores were rounded up to the nearest whole number if the total portfolio score ended in .5.
- Occasionally, portfolios are submitted that do not meet submission requirements and result in a condition code for one or more rubrics. A condition code explaining the reason a rubric is deemed unscorable is reported to the candidate. No portfolios with condition codes were included in these reports and analyses.

Means and distributions of total scores are not provided for fields with fewer than 10 portfolios. Fields with fewer than 10 portfolios are omitted from the rubric-level distribution reporting tables. Note that estimates based on sample sizes below 100 may be unstable and should be interpreted with caution.

Table B-1: Average candidate performance for 15-rubric content fields

	N	Total Score Mean	Planning					Instruction					Assessment					Mean by Task		
			P01	P02	P03	P04	P05	I06	I07	I08	I09	I10	A11	A12	A13	A14	A15	P	I	A
All 15-Rubric Handbooks	49,123	44.16	3.0	3.0	3.1	2.9	2.9	3.1	3.0	2.9	2.9	2.7	2.9	3.2	2.7	2.9	2.9	14.9	14.5	14.7
Agricultural Education	267	47.10	3.4	3.2	3.2	3.1	3.2	3.1	3.1	3.1	3.4	3.0	3.2	3.3	2.9	2.9	3.1	16.1	15.7	15.3
Business Education	200	42.34	3.0	2.8	2.8	2.9	3.0	3.1	2.9	2.7	2.9	2.5	2.9	3.2	2.3	2.7	2.7	14.4	14.1	13.8
Early Childhood	4,195	41.49	2.9	2.8	2.8	3.0	2.9	3.0	3.0	2.9	2.3	2.7	2.7	2.6	2.4	2.9	2.8	14.4	13.8	13.3
Elementary Education: Literacy with Mathematics Task 4	10,288	43.93	2.9	2.9	3.1	2.9	2.8	3.0	2.9	2.9	2.9	2.8	3.0	3.3	2.8	2.8	3.1	14.5	14.4	15.0
Elementary Education: Mathematics with Literacy Task 4	571	46.75	3.2	3.1	3.3	3.1	3.0	3.1	3.1	3.0	3.1	3.0	3.3	3.4	3.0	3.1	3.2	15.6	15.3	15.9

	N	Total Score Mean	Planning					Instruction					Assessment					Mean by Task		
			P01	P02	P03	P04	P05	I06	I07	I08	I09	I10	A11	A12	A13	A14	A15	P	I	A
Elementary Literacy	4,055	44.53	3.0	2.9	3.1	3.0	2.8	3.0	2.9	2.9	3.0	2.7	3.0	3.4	2.9	2.8	3.1	14.8	14.5	15.2
Elementary Mathematics	4,006	45.25	3.2	3.0	3.2	2.9	2.9	3.1	3.0	2.9	3.1	2.8	3.2	3.3	2.9	2.8	3.0	15.2	14.8	15.3
English as an Additional Language	632	47.85	3.4	3.2	3.4	3.3	3.3	3.3	3.2	3.1	2.9	3.0	3.2	3.3	2.9	3.1	3.2	16.6	15.5	15.7
Family and Consumer Sciences	127	44.49	3.2	3.1	3.1	3.1	3.0	3.1	2.9	2.9	2.9	2.9	3.0	3.0	2.7	2.8	2.9	15.5	14.7	14.3
Health Education	148	37.84	2.7	2.5	2.5	2.4	2.7	3.0	2.6	2.5	2.3	2.4	2.6	2.8	2.2	2.2	2.3	12.8	12.9	12.2
K–12 Performing Arts	2,018	45.91	3.1	3.1	3.3	3.0	3.1	3.1	3.0	2.8	3.0	2.8	3.2	3.3	2.9	3.0	3.2	15.6	14.7	15.6
Library Specialist	74	47.07	3.5	3.5	3.5	3.1	3.3	3.2	3.0	2.9	3.1	3.1	3.2	3.0	2.8	2.8	3.1	16.9	15.3	14.9
Middle Childhood English-Language Arts	506	47.67	3.3	3.3	3.4	3.2	3.3	3.3	3.1	3.2	3.0	2.9	3.1	3.5	2.9	2.9	3.2	16.5	15.5	15.7
Middle Childhood History/Social Studies	451	45.04	3.1	3.0	3.3	2.9	2.9	3.1	3.0	2.9	2.8	2.7	3.1	3.5	2.9	2.8	3.0	15.3	14.4	15.3
Middle Childhood Mathematics	710	44.16	3.1	2.8	3.2	2.8	2.9	3.1	2.8	2.9	3.1	2.6	3.1	3.4	2.8	2.8	2.8	14.8	14.4	14.9
Middle Childhood Science	430	45.44	3.2	3.1	3.3	3.1	3.1	3.1	2.9	2.9	2.7	2.7	3.1	3.3	2.9	2.9	3.0	15.8	14.4	15.2
Physical Education	1,527	40.16	2.9	2.9	2.7	2.6	2.6	2.8	2.8	3.0	2.9	2.5	2.5	2.9	2.3	2.4	2.5	13.7	13.9	12.5
Secondary English-Language Arts	3,443	46.38	3.3	3.1	3.2	3.0	3.1	3.1	3.1	3.0	3.0	2.9	3.2	3.4	2.8	3.0	3.1	15.8	15.1	15.5
Secondary History/Social Studies	3,324	44.85	3.1	3.0	3.0	2.9	3.0	3.1	3.0	3.0	2.9	2.8	3.1	3.4	2.8	2.8	3.0	15.0	14.8	15.1
Secondary Mathematics	2,144	40.21	2.8	2.4	2.8	2.4	2.6	3.0	2.5	2.6	2.8	2.5	2.6	3.3	2.6	2.6	2.6	13.1	13.4	13.7
Secondary Science	2,531	42.42	3.0	2.7	3.0	2.8	2.9	3.1	2.8	2.8	2.6	2.5	2.8	3.2	2.6	2.9	2.8	14.5	13.7	14.2
Special Education	6,370	44.55	3.0	3.1	3.1	3.1	2.9	3.2	3.1	3.0	3.1	2.7	2.7	3.1	2.7	3.0	2.7	15.2	15.1	14.2
Technology and Engineering Education	122	45.18	3.2	3.0	3.1	3.1	2.9	3.1	3.0	2.8	3.0	2.8	3.1	3.3	2.8	2.9	3.1	15.2	14.7	15.2
Visual Arts	984	48.83	3.7	3.4	3.5	3.3	3.5	3.3	3.2	3.1	3.2	2.9	3.3	3.5	2.8	2.9	3.2	17.4	15.7	15.7

Table B-2: Average candidate performance for 13-rubric content fields

	N	Total Score Mean	Planning					Instruction					Assessment					Mean by Task		
			P01	P02	P03	P04	P05	I06	I07	I08	I09	I10	A11	A12	A13	A14	A15	P	I	A
All 13-Rubric Handbooks	906	35.70	3.1	3.0	3.0		3.0	3.1	2.7	2.5	2.0	2.5	2.7	2.9	2.5		2.7	12.1	12.8	10.9
Classical Languages	15	32.93	2.9	2.8	3.1		2.7	3.0	2.3	2.3	1.9	2.4	2.0	2.5	2.5		2.6	11.5	11.9	9.5
World Language	891	35.75	3.1	3.0	3.0		3.0	3.1	2.7	2.5	2.0	2.5	2.7	3.0	2.5		2.7	12.1	12.8	10.9

Table B-3: Average candidate performance for 18-rubric content fields

	N	Total Score Mean	Planning					Instruction					Assessment					Mathematics			Mean by Task		
			P01	P02	P03	P04	P05	I06	I07	I08	I09	I10	A11	A12	A13	A14	A15	M19	M20	M21	P	I	A
All 18-Rubric Handbooks	10,714	52.64	2.9	2.9	3.1	2.9	2.8	3.0	2.9	2.9	2.9	2.8	3.0	3.3	2.9	2.9	3.1	2.8	2.9	2.7	14.6	14.5	15.1
Elementary Education: Literacy with Mathematics Task 4	10,150	52.45	2.9	2.9	3.1	2.9	2.8	3.0	2.9	2.9	2.9	2.8	3.0	3.3	2.9	2.9	3.1	2.8	2.9	2.7	14.5	14.4	15.1
Elementary Education: Mathematics with Literacy Task 4	564	55.93	3.2	3.1	3.3	3.1	3.0	3.1	3.1	3.0	3.1	3.0	3.3	3.4	3.0	3.1	3.2	2.9	3.0	3.2	15.6	15.3	15.9

Appendix C: Score Distributions by Content Field

The following tables present the mean scores and distribution of total scores across 15-, 13-, and 18-rubric national content fields.

Table C-1: Mean scores and distribution of total scores across 15-rubric national content fields

			Distribution of Total Score (%)									
	N	Mean Score	< 35	35	36	37	38	39	40	41	42	> 42
All 15-Rubric Handbooks	49,123	44.16	7	2	2	3	2	3	4	5	5	67
Agricultural Education	267	47.10	6	2	0	1		1	1	1	3	84
Business Education	200	42.34	7	2	1	4	5	3	9	9	10	52
Early Childhood	4,195	41.49	7	3	4	5	4	5	8	8	8	47
Elementary Education: Literacy with Mathematics Task 4	10,288	43.93	7	2	2	2	2	3	3	4	6	69
Elementary Education: Mathematics with Literacy Task 4	571	46.75	3	1	1	1	0	1	1	2	2	88
Elementary Literacy	4,055	44.53	6	2	2	2	1	2	3	4	5	72
Elementary Mathematics	4,006	45.25	3	1	1	2	1	2	3	4	4	77
English as an Additional Language	632	47.85	5	1	1	2	0	0	2	2	3	82
Family and Consumer Sciences	127	44.49	4	2	5	1	3	2	2	7	6	69
Health Education	148	37.84	34	7	5	4	7	2	2	3	4	32
K–12 Performing Arts	2,018	45.91	6	1	1	3	1	1	3	4	5	74
Library Specialist	74	47.07	9	3				1	3	3	5	76
Middle Childhood English-Language Arts	506	47.67	7	1	2	2	1	2	2	3	3	78
Middle Childhood History/Social Studies	451	45.04	10	1	2	3	1	1	4	3	4	71
Middle Childhood Mathematics	710	44.16	7	2	3	3	2	3	4	6	5	64
Middle Childhood Science	430	45.44	7	1	1	2	1	3	2	4	5	73
Physical Education	1,527	40.16	17	5	4	5	4	4	8	9	6	37
Secondary English-Language Arts	3,443	46.38	3	1	1	1	1	2	3	3	3	82
Secondary History/Social Studies	3,324	44.85	5	2	2	2	1	2	3	4	5	75
Secondary Mathematics	2,144	40.21	17	5	6	5	4	4	7	8	7	37
Secondary Science	2,531	42.42	13	3	4	4	3	3	4	6	6	54
Special Education	6,370	44.55	8	1	2	2	2	2	4	4	5	68
Technology and Engineering Education	122	45.18	7	3	1	2	1	3	3	1	8	70
Visual Arts	984	48.83	2	1	1	1	1	1	3	2	3	86

Table C-2: Mean scores and distribution of total scores across 13-rubric national content fields

			Distribution of Total Score (%)							
	N	Mean Score	< 30	30	31	32	33	34	35	> 35
All 13-Rubric Handbooks	906	35.70	18	4	6	4	5	6	5	53
Classical Languages	15	32.93	20	13	7			13	7	40
World Language	891	35.75	18	3	6	4	5	6	4	53

Table C-3: Mean scores and distribution of total scores across 18-rubric national content fields

			Distribution of Total Score (%)											
	N	Mean Score	< 40	40	41	42	43	44	45	46	47	48	49	> 49
All 18-Rubric Handbooks	10,714	52.64	5	1	2	2	2	2	1	2	2	3	3	75
Elementary Education: Literacy with Mathematics Task 4	10,150	52.45	5	1	2	2	2	2	1	2	2	3	4	74
Elementary Education: Mathematics with Literacy Task 4	564	55.93	2	1	1	1	0	1	0	1	1	1	1	91

Appendix D: Portfolios Represented by State

The following table shows all states that submitted edTPA portfolios during the 2019 administrative year and the approximate percentage of the total sample that each state contributed.

Table D-1: Approximate percentage of the total sample each state contributed

State	% of Total Sample
AL	3.7%
AR	<1%
AZ	<1%
CA	7.9%
CO	<1%
CT	2.9%
DE	<1%
GA	8.5%
HI	<1%
IA	<1%
IL	8.4%
IN	1.1%
KY	<1%
MD	1.3%
MN	5.3%
NC	5.3%
NE	<1%

State	% of Total Sample
NJ	5.7%
NY	15.5%
OH	6.8%
OK	<1%
OR	3.3%
PA	<1%
SC	<1%
TN	4.4%
TX	<1%
UT	<1%
VA	<1%
WA	6%
WI	<1%
WV	<1%
WY	<1%

Appendix E: Consequential Use by Content Field

The following tables present a comparison of average scores and standard deviations of portfolios from all states, from states without consequential policy, and from states with consequential policy for all 15-, 13-, and 18-rubric handbooks.

Table E-1: Average scores and standard deviations from all states, from states without policy, and from states with policy for 15-rubric content fields

		All			Non-Policy			Policy States		
	Field	N	Mean	Std Deviation	N	Mean	Std Deviation	N	Mean	Std Deviation
15 Rubrics	Agricultural Education	267	47.10	6.96	47	42.74	8.38	220	48.03	6.26
	Business Education	200	42.34	4.55	14	41.43	4.89	186	42.41	4.53
	Early Childhood	4,195	41.49	4.58	1,600	41.08	5.03	2,595	41.75	4.25
	Educational Technology Specialist	3								
	Elementary Education: Literacy with Mathematics Task 4	10,288	43.93	5.50	1,126	41.47	6.48	9,162	44.23	5.29
	Elementary Education: Mathematics with Literacy Task 4	571	46.75	5.32	47	43.28	5.22	524	47.06	5.22
	Elementary Literacy	4,055	44.53	5.63	309	42.26	6.59	3,746	44.72	5.50
	Elementary Mathematics	4,006	45.25	5.22	263	43.10	6.16	3,743	45.40	5.11
	English as an Additional Language	632	47.85	7.22	65	46.05	8.46	567	48.06	7.04
	Family and Consumer Sciences	127	44.49	5.48	22	42.82	6.69	105	44.84	5.16
	Health Education	148	37.84	7.91	34	37.71	7.75	114	37.89	8.00
	K–12 Performing Arts	2,018	45.91	6.48	273	44.30	6.97	1,745	46.16	6.37
	Library Specialist	74	47.07	8.26	2	41.50	9.19	72	47.22	8.25
	Literacy Specialist	7								
	Middle Childhood English-Language Arts	506	47.67	8.00	163	45.76	8.39	343	48.57	7.66
	Middle Childhood History/Social Studies	451	45.04	7.04	119	44.73	8.20	332	45.15	6.59
	Middle Childhood Mathematics	710	44.16	6.28	186	43.99	6.39	524	44.22	6.24
	Middle Childhood Science	430	45.44	6.66	113	45.13	6.85	317	45.55	6.60
	Physical Education	1,527	40.16	5.83	106	37.91	6.33	1,421	40.33	5.76
	Secondary English-Language Arts	3,443	46.38	5.67	492	44.87	6.28	2,951	46.63	5.52
	Secondary History/Social Studies	3,324	44.85	5.39	468	42.54	6.17	2,856	45.23	5.15
	Secondary Mathematics	2,144	40.21	5.31	262	38.71	5.66	1,882	40.42	5.23
	Secondary Science	2,531	42.42	6.43	275	39.99	6.92	2,256	42.72	6.30
	Special Education	6,370	44.55	6.45	1,146	42.81	6.85	5,224	44.93	6.29
	Technology and Engineering Education	122	45.18	6.64	14	45.50	9.45	108	45.14	6.25
	Visual Arts	984	48.83	6.83	162	45.81	7.09	822	49.43	6.62
	All 15-Rubric Handbooks	49,123	44.16	6.10	7,311	42.30	6.59	41,822	44.49	5.95

Table E-2: Average scores and standard deviations from all states, from states without policy, and from states with policy for 13-rubric content fields

		All			Non-Policy			Policy States		
	Field	N	Mean	Std Deviation	N	Mean	Std Deviation	N	Mean	Std Deviation
13 Rubrics	Classical Languages	15	32.93	5.01	4	30.50	7.77	11	33.82	3.71
	World Language	891	35.75	6.30	119	34.20	6.03	772	35.99	6.31
	All 13-Rubric Handbooks	906	35.70	6.29	123	34.08	6.09	783	35.96	6.28

Table E-3: Average scores and standard deviations from all states, from states without policy, and from states with policy for 18-rubric content fields

		All			Non-Policy			Policy States		
	Field	N	Mean	Std Deviation	N	Mean	Std Deviation	N	Mean	Std Deviation
18 Rubrics	Elementary Education: Literacy with Mathematics Task 4	10,150	52.45	6.64	1,099	49.54	7.72	9,051	52.81	6.40
	Elementary Education: Mathematics with Literacy Task 4	564	55.93	6.36	45	51.42	6.59	519	56.32	6.19
	All 18-Rubric Handbooks	10,714	52.64	6.67	1,144	49.61	7.69	9,570	53.00	6.44

Appendix F: ANOVAs and Post-Hoc Analyses

Each ANOVA was run using the general linear model because it is more robust to unbalanced group sizes. Within an ANOVA table, a Sig* indicates the groups do not have homogeneous variances and the significance comes from a Welch's ANOVA, which better controls the Type I error rate in the case of heterogeneous variance.

Note: Post-hoc analyses presented do not include portfolios that are categorized as "other" or "unidentified" or which have a sample size less than 100, with the exception of Gender, where they were included. Due to unequal sample sizes, all comparisons should be interpreted with caution.

Table F-1: School Locale

ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	4	4032.067	1008.017	28.58	<.0001
Error	41817	1474639.355	35.264		
Corrected Total	41821	1478671.422			

Post-Hoc Analyses

School Locale	Difference Between Means	Sig. at 0.05
Suburb – City	0.11921	
Suburb – Rural	0.35998	***
Suburb – Town	0.79811	***
City – Suburb	-0.11921	
City – Rural	0.24077	***
City – Town	0.67890	***
Rural – Suburb	-0.35998	***
Rural – City	-0.24077	***
Rural – Town	0.43813	***
Town – Suburb	-0.79811	***
Town – City	-0.67890	***
Town – Rural	-0.43813	***

*** Comparisons significant at the 0.05 level

Table F-2: Percentage ELL

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	5	1467.964	293.593	8.31	<.0001
Error	41816	1477203.457	35.326		
Corrected Total	41821	1478671.422			

Percent of English Language Learners	Difference Between Means	Sig. at 0.05
76–100% - 26–50%	0.07923	
76–100% - 10–25%	0.12010	
76–100% - <10%	0.24240	
76–100% - 51–75%	0.57070	
26–50% - 76–100%	-0.07923	
26–50% - 10–25%	0.04087	
26–50% - <10%	0.16317	
26–50% - 51–75%	0.49147	
10–25% - 76–100%	-0.12010	
10–25% - 26–50%	-0.04087	
10–25% - <10%	0.12230	
10–25% - 51–75%	0.45060	
<10% - 76–100%	-0.24240	
<10% - 26–50%	-0.16317	
<10% - 10–25%	-0.12230	
<10% - 51–75%	0.32830	
51–75% - 76–100%	-0.57070	
51–75% - 26–50%	-0.49147	
51–75% - 10–25%	-0.45060	
51–75% - <10%	-0.32830	

Table F-3: Percentage Free and Reduced-Price Lunch Elibability

Source	DF	Sum of Squares	Mean Square	F Value	Sig
Model	5	2045.727	409.145	11.59	<.0001
Error	41816	1476625.695	35.312		
Corrected Total	41821	1478671.422			

Percent Qualified for Free and Reduced-Price Lunch	Difference Between Means	Sig. at 0.05
26–50% - 10–25%	0.02541	
26–50% - 51–75%	0.29458	
26–50% - 76–100%	0.39534	***
26–50% - <10%	0.41807	***
10–25% - 26–50%	-0.02541	
10–25% - 51–75%	0.26917	
10–25% - 76–100%	0.36993	***
10–25% - <10%	0.39266	***
51–75% - 26–50%	-0.29458	
51–75% - 10–25%	-0.26917	
51–75% - 76–100%	0.10076	
51–75% - <10%	0.12349	
76–100% - 26–50%	-0.39534	***
76–100% - 10–25%	-0.36993	***
76–100% - 51–75%	-0.10076	
76–100% - <10%	0.02273	
<10% - 26–50%	-0.41807	***
<10% - 10–25%	-0.39266	***
<10% - 51–75%	-0.12349	
<10% - 76–100%	-0.02273	

Table F-4: Percentage SPED Eligibility

ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	5	1134.233	226.847	6.42	<.0001
Error	41816	1477537.189	35.334		
Corrected Total	41821	1478671.422			

Percent of Students Eligible for Special Education	Difference Between Means	Sig. at 0.05
<10% - 76–100%	0.04042	
<10% - 26–50%	0.14014	
<10% - 10–25%	0.14901	
<10% - 51–75%	0.75948	***
76–100% - <10%	-0.04042	
76–100% - 26–50%	0.09972	
76–100% - 10–25%	0.10860	
76–100% - 51–75%	0.71906	
26–50% - <10%	-0.14014	
26–50% - 76–100%	-0.09972	
26–50% - 10–25%	0.00887	
26–50% - 51–75%	0.61934	
10–25% - <10%	-0.14901	
10–25% - 76–100%	-0.10860	
10–25% - 26–50%	-0.00887	
10–25% - 51–75%	0.61047	
51–75% - <10%	-0.75948	***
51–75% - 76–100%	-0.71906	
51–75% - 26–50%	-0.61934	
51–75% - 10–25%	-0.61047	

Table F-5: Placement Length

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	6	3937.768	656.295	18.61	<.0001
Error	41815	1474733.654	35.268		
Corrected Total	41821	1478671.422			

Length of Placement	Difference Between Means	Sig. at 0.05
21–30 - 31–40	0.26363	
21–30 - 11–20	0.46965	***
21–30 - 1–10	0.48762	***
21–30 - 41+	0.59013	***
21–30 - 0	2.28651	***
31–40 - 21–30	-0.26363	
31–40 - 11–20	0.20602	
31–40 - 1–10	0.22399	
31–40 - 41+	0.32651	
31–40 - 0	2.02288	***
11–20 - 21–30	-0.46965	***
11–20 - 31–40	-0.20602	
11–20 - 1–10	0.01797	
11–20 - 41+	0.12049	
11–20 - 0	1.81686	***
1–10 - 21–30	-0.48762	***
1–10 - 31–40	-0.22399	
1–10 - 11–20	-0.01797	
1–10 - 41+	0.10251	
1–10 - 0	1.79889	***
41+ - 21–30	-0.59013	***
41+ - 31–40	-0.32651	
41+ - 11–20	-0.12049	
41+ - 1–10	-0.10251	
41+ - 0	1.69637	***
0 - 21–30	-2.28651	***

Length of Placement	Difference Between Means	Sig. at 0.05
0 - 31–40	-2.02288	***
0 - 11–20	-1.81686	***
0 - 1–10	-1.79889	***
0 - 41+	-1.69637	***

Table F-6: Preparation Route

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	5	6621.267	1324.253	37.62	<.0001
Error	41816	1472050.155	35.203		
Corrected Total	41821	1478671.422			

Teacher Preparation Route	Difference Between Means	Sig. at 0.05
Traditional EPP - Teach for America	0.02478	
Traditional EPP - Alt - university or college-based program	0.89690	***
Traditional EPP - Alt - district provider	1.31391	***
Traditional EPP - Other	2.09290	***
Teach for America - Traditional EPP	-0.02478	
Teach for America - Alt - university or college-based program	0.87212	***
Teach for America - Alt - district provider	1.28913	***
Teach for America - Other	2.06812	***
Alt - university or college-based program - Traditional EPP	-0.89690	***
Alt - university or college-based program - Teach for America	-0.87212	***
Alt - university or college-based program - Alt - district provider	0.41701	
Alt - university or college-based program - Other	1.19600	***
Alt - district provider - Traditional EPP	-1.31391	***
Alt - district provider - Teach for America	-1.28913	***

Teacher Preparation Route	Difference Between Means	Sig. at 0.05
Alt - district provider - Alt - university or college-based program	-0.41701	
Alt - district provider - Other	0.77899	
Other - Traditional EPP	-2.09290	***
Other - Teach for America	-2.06812	***
Other - Alt - university or college-based program	-1.19600	***
Other - Alt - district provider	-0.77899	

Table F-7: Ethnicity

ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	7	10586.552	1512.365	43.08	<.0001
Error	41814	1468084.870	35.110		
Corrected Total	41821	1478671.422			

Post-Hoc Analyses

Ethnicity	N by Group	Difference Between Means	Sig. at 0.05
Asian - White	1,744; 30,910	-0.06047	
Asian - Hispanic	1,744; 3,508	0.20511	
Asian - Multiracial	1,744; 1,390	0.27492	
Asian - Native American	1,744; 141	1.64090	***
Asian - Black	1,744; 2,726	1.88351	***
Black - White	2,726;30,910	-1.94398	***
Black - Asian	2,726; 1,744	-1.88351	***
Black - Hispanic	2,726; 3,505	-1.67839	***
Black - Multiracial	2,726; 1,390	-1.60859	***
Black - Native American	2,726; 141	-0.24260	
Hispanic - White	3,508; 30,910	-0.26558	
Hispanic - Asian	3,508; 1,744	-0.20511	

Ethnicity	N by Group	Difference Between Means	Sig. at 0.05
Hispanic - Multiracial	3,508; 1,390	0.06981	
Hispanic - Native American	3,508; 141	1.43579	
Hispanic - Black	3,508; 2,726	1.67839	***
Multiracial - White	1,390; 30,910	-0.33539	
Multiracial - Asian	1,390; 1,744	-0.27492	
Multiracial - Hispanic	1,390; 3,508	-0.06981	
Multiracial - Native American	1,390; 141	1.36598	
Multiracial - Black	1,390; 2,726	1.60859	***
Native American - White	141; 30,910	-1.70137	***
Native American - Asian	141; 1,744	-1.64090	***
Native American - Hispanic	141; 3,508	-1.43579	
Native American - Multiracial	141; 1,390	-1.36598	
Native American - Black	141; 2,726	0.24260	
White - Asian	30,910; 1,744	0.06047	
White - Hispanic	30,910; 3,508	0.26558	
White - Multiracial	30,910; 1,390	0.33539	
White - Native American	30,910; 141	1.70137	***
White - Black	30,910; 2,726	1.94398	***

*** Comparisons significant at the 0.05 level

Table F-8 Primary Language

ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	2	545.176	272.588	7.71	0.0004
Error	41819	1478126.246	35.346		
Corrected Total	41821	1478671.422			

Post-Hoc Analyses

Is English Your Primary Language?	Difference Between Means	Sig. at 0.05
Yes - No	0.62469	***
No - Yes	-0.62469	***

Table F-9: Gender

ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	1	17091.127	17091.127	488.94	<.0001
Error	41512	1451070.686	34.955		
Corrected Total	41513	1468161.813			

Post-Hoc Analyses

Gender	Difference Between Means	Sig. at 0.05
Male - Female	-1.55199	***
Female - Male	1.55199	***

*** Comparisons significant at the 0.05 level

Table F-10: Education Level

ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Sig*
Model	3	1046.983	348.994	9.88	<.0001
Error	41818	1477624.439	35.335		
Corrected Total	41821	1478671.422			

Post-Hoc Analyses

Education Level	Difference Between Means	Sig. at 0.05
Doctorate - Bachelor's/Bachelor's plus	0.00280	
Doctorate - High school/Some college	0.02441	
Doctorate - Master's/Master's plus credits	0.55424	
Bachelor's/Bachelor's plus - Doctorate	-0.00280	
Bachelor's/Bachelor's plus - High school/Some college	0.02161	
Bachelor's/Bachelor's plus - Master's/Master's plus credits	0.55144	***
High school/Some college - Doctorate	-0.02441	
High school/Some college - Bachelor's/Bachelor's plus	-0.02161	
High school/Some college - Master's/Master's plus credits	0.52983	***
Master's/Master's plus credits - Doctorate	-0.55424	
Master's/Master's plus credits - Bachelor's/Bachelor's plus	-0.55144	***
Master's/Master's plus credits - High school/Some college	-0.52983	***

*** Comparisons significant at the 0.05 level

Citations

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.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-raters' performance on TOEFL testing. *ETS Research Report Series*, *2004*(1), i–38.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal*, *47*(1), 218–251.

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: Comments on Kappa, Cohen's Kappa, Scott's π, and Aickin's α. *Understanding Statistics, 2*(3), 205–219.

Koo, T. K., & Lee, M. Y., (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Pecheone, R. L., &, Whittaker, A. (2016). Well prepared teachers inspire student learning. *Kappan*, *97*(7), 8–13.

Powers, D. E. (2000). Computing reader agreement for the GRE Writing Assessment. Princeton, NJ: Educational Testing Service.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

SAS (2017). SAS/STAT® 14.3 User's Guide, SAS Institute Inc., Cary, NC, USA.

Stanford Center for Assessment, Learning and Equity (SCALE). (2013). edTPA Field test: Summary report. Palo Alto, CA: Author.

Stanford Center for Assessment, Learning and Equity (SCALE). (2015). Educative assessment and meaningful support: 2014 edTPA Administrative Report. Palo Alto, CA: Author.

Stanford Center for Assessment, Learning and Equity (SCALE). (2016). Educative assessment and meaningful support: 2015 edTPA Administrative Report. Palo Alto, CA: Author.

Stanford Center for Assessment, Learning and Equity (SCALE). (2017). Educative assessment and meaningful support: 2016 edTPA Administrative Report. Palo Alto, CA: Author.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. Osborne (*Ed.*), Best Practices in Quantitative Methods, (pp. 29–49). SAGE Publications, Inc.: https://www.doi.org/10.4135/9781412995627

Tanguay, C. L., Many, J. E., Ariail, M., Bhatnagar, R., & Emerson, J. (2019). An Organic Model for edTPA Exploration and Implementation. In K. Winter, H. Pinter, & M. Watson (Eds.), *Performance-Based Assessment in 21st Century Teacher Education* (pp. 42–80). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-8353-0.ch003

Whittaker, A., Pecheone, R., and Stansbury, K. (2018). Fulfilling our educative mission: A response to edTPA critique. *Education Policy Analysis Archives*, *26*(30), 1–16.