Released October 2016

EDUCATIVE ASSESSMENT & MEANINGFUL SUPPORT

2015 edTPA Administrative Report

edTPA®

Preface and Acknowledgements

edTPA is a performance assessment for pre-service teacher candidates, which was developed and field tested beginning in 2009 and has been used operationally since September 2013. This report presents analyses based on teacher candidate performance from **January 1st to December 31st, 2015,** and complements prior reports that have reviewed the development of the assessment, as previously described in detail in both the 2013 edTPA Field Test Summary Report, and the 2014 edTPA Annual Administrative Report.

This administrative report was authored by: Raymond L. Pecheone, Executive Director, Stanford Center for Assessment, Learning and Equity (SCALE); Andrea Whittaker, Director, Teacher Performance Assessment, SCALE; and Heather Klesch, Director, Educator Solutions for Licensing and Learning, Evaluation Systems group of Pearson. We thank Irena Nayfeld, post doctoral scholar (SCALE) and Ben Shear, Stanford Doctoral Candidate for their contributions to the analyses in 2015 and 2016.

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.

The structure of this report and its contents are highly similar to the 2014 Administrative report and represent a framework we will use annually to present edTPA candidate performance data and reliability and validity evidence. Some sections that appeared in the 2014 report are summarized here, however the 2014 report may be referenced for additional information. The analyses presented this year replicate those conducted for the 2014 Administrative Report and were reviewed by technical committee members and advisors again this year. See Appendix I for a complete list of TAC members. We are grateful to them for their advice and recommendations, which continue to strengthen the analyses edTPA and inform an ongoing research agenda.

We also are grateful to the early funders of the research and development process, including the Ford Foundation, the MetLife Foundation, the Morgan Family Foundation, the Stuart Foundation and the Hewlett Foundation. We continue to be grateful for the input and critique of the hundreds of teachers and teacher educators who participated in past handbook and support resource development as design team members, content validation participants, bias and sensitivity reviewers, scorers, trainers, and supervisors as well as Educator Preparation Program (EPP) faculty who have piloted, field tested, and implemented edTPA since 2009.

As developers of edTPA, 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 Evaluation Systems group of Pearson are pleased to release the 2015 Administrative Report. This report presents all candidate performance data from the 27,000+ candidates who participated in edTPA during the second full operational year (January 1 to December 31, 2015), and 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. As in 2014, all analyses and results have been informed and reviewed by a technical advisory committee of nationally recognized psychometricians, and meet the technical standards for licensure assessments set forth by AERA, APA, & NCME (2014).

SCALE and AACTE commend the nearly 700 educator preparation programs in 38 states that contributed to the development and field testing[1] 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[2] 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 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 teachers entering the profession**. It provides a measure of teacher candidates' readiness to teach that can inform licensure, accreditation decisions, and program completion. Most 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. 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 an authentic cycle of teaching aimed at subject-specific student learning goals, using 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

[1] See the edTPA Summary Report 2013 for a complete description of edTPA development, field testing and candidate performance prior to operational use.

[2] See the 2014 edTPA Annual Administrative Report for additional information on the edTPA development, and for operational candidate performance from the 2014 administration year.

three tasks are candidates' abilities to develop their students' academic language and to justify and analyze their own teaching practices.

All 27 edTPA handbooks share approximately 80% of their design, assessing pedagogical constructs that underlie the integrated cycle of planning, instruction, and assessment. 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 assessment requires that the central focus of their learning segment supports students' development of conceptual understanding, procedural fluency, and problem solving/reasoning skills of a standards-based topic, that their lesson design includes mathematics-pertinent language demands and supports, and that assessments provide opportunities for students to demonstrate development of mathematics concepts and reasoning skills.

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 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 (Pecheone & Whittaker, 2016). In addition, the new 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 about 9,100 faculty from approximately 700 teacher preparation programs who have downloaded the program's 165+ implementation resources over 670,000 times. 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 2,500 qualified teachers and teacher educators now serve as scorer trainers, supervisors, or scorers. Scorers must be P-12 teachers or teacher preparation faculty with significant pedagogical content knowledge in the field in which they score, as well as experience working as instructors or mentors for novice teachers (e.g., NBTPS teachers). In the 2015 administration year (January 1st, 2015 – December 31st, 2015), scorer recruitment goals targeted a balance of approximately 50% teacher educators and 50% practicing classroom teachers; of these, 32% of practicing classroom teachers and 20% of the qualified scoring pool are National Board certified teachers. 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.

Scorer reliability was evaluated using several different statistical tests. In a random sample of 2,617 portfolios double-scored independently by two scorers, the scorers assigned either the same or adjacent scores (total agreement) in approximately 95% of all cases. Kappa n agreement rates reveal that scorers tend to assign scores within +/- 1 and rarely assign scores

that differ by more than 1 point (overall kappa n reliability = .89). Internal consistency of the 15 rubrics, or items, was evaluated using Cronbach's alpha (.91) and a latent trait IRT partial credit model that produced a reliability estimate of (0.910). As in 2014, all reliability coefficients indicate a high degree of internal consistency of rubrics to the measured construct (readiness to teach). These results are consistent with 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.

Validity Evidence

edTPA was developed as an authentic, subject-specific, performance-based support and assessment system of a candidate's readiness to teach. Following the validity guidelines presented in the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), this report defines the constructs assessed by edTPA and presents evidence that examines its use and interpretations. The validity section reviews sources of validity evidence for edTPA; these include the empirical research and theory on which the development was based, the design process and content 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. Results from a Confirmatory Factor Analyses (CFA) and a polytomous item response theory (IRT) model provide empirical support for the edTPA constructs of planning, instruction, and assessment.

Candidate Performance

This report presents performance data from 27,759 submissions: average scores and distributions overall by task and by rubric for the entire sample, as well as for each of the 27 content fields. The total score, computed as an aggregation of scores on a 5-point scale across 15 rubrics, ranges from 5 to 75 total points. The average edTPA score across 27,172 portfolios from fields with 15-rubric handbooks (including the first 15 rubrics of Elementary Education) was 44.2, with a standard deviation of 7.42. Performance by task is an aggregation of scores on the 5 rubrics that make up each task; these range from 5 to 25 points for each task. Over a number of field trials and in operational use, a consistent candidate performance across edTPA teaching tasks has emerged: candidates performed most strongly on the planning task (M = 15.3), followed by the instruction task (M = 14.7) and the assessment task (M = 14.2). This conforms to other studies that have found that learning to evaluate and respond to students' learning and provide meaningful feedback is one of the more challenging elements of teaching (Black & William, 1998; Otero, 2006; Siegel & Wissehr, 2011).

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. The overall mean score for 15-rubric fields for all candidates in states with consequential policy was 44.53, (N of 21,452). Based on the national recommended professional performance standard of 42 (note that to date no state has a cut score of 42), the pass rate for all candidates who submitted an edTPA portfolio in 2015 was 71% across all states, and 72% in states using the assessment consequentially. Note that cut scores vary by state as do passing rates, and to date 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 profession.

When submitting an edTPA portfolio for official scoring, the candidate is asked to provide demographic information in several categories: gender, ethnicity, teaching placement context, 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 of 21,452). These analyses revealed that all demographic variables taken

together explained approximately 4% of the total variance in edTPA scores. Differences by racial /ethnic group were small, women generally scored more highly than men, and suburban teachers on average scored more highly than teachers in other teaching contexts. Performance differences were found between African American and White candidates, with differences in mean performance at about one half of a standard deviation. In addition, White and Hispanic candidates had comparable performance, as did those indicating Other for ethnicity, and those that declined to answer. Small sample sizes for some groups and differences in group sizes prevent strong generalizations.

edTPA is committed to providing equal opportunity for all teacher candidates and will continue 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.

Next Steps for Research

The input of the edTPA National Technical Advisory Committee guided the analyses and interpretations presented in this report; their recommendations and feedback are reflected throughout. Additional research recommendations from the TAC and a recently convened group of teacher education scholars will continue to inform ongoing studies of consequential impact, predictive validity, and other areas of research interest.

Conclusion

As with the Field Test data and those of the 2014 Administrative Report, data presented here are consistent with the technical standards of APA, AERA and NCME (2014) and support the use of edTPA to grant an initial license to pre-service teacher candidates as well as to inform state and national accreditation. The reporting of performance of all candidates who submitted edTPA portfolios in 2015 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 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 formative 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.

Introduction

By the Profession, for the Profession

Based upon a 25-year history of assessment development led by Raymond Pecheone and Linda Darling-Hammond, edTPA draws on the experiences in developing performance-based assessments including the National Board for Professional Teaching Standards' (NBPTS) assessments of accomplished veteran teachers, the Interstate Teacher Assessment and Support Consortium (InTASC) Portfolio, and the Performance Assessment for California Teachers (PACT). These portfolio-based designs have stood the test of time and consistently reveal key features of effective teaching. After more than four years of development and analysis, including two years of field testing with more than 12,000 teacher candidates, edTPA was launched operationally in September 2013 as a performance-based assessment to measure the classroom practice of pre-service teacher candidates – to ensure they are ready to teach on day one. The assessment was developed by faculty and staff at Stanford University with leadership by the American Association of Colleges for Teacher Education (AACTE), subject-specific design teams comprised of teachers and teacher educators who are subject-matter experts, and substantive advice and feedback from educators nationwide. More than 1,000 educators from 29 states and the District of Columbia and more than 430 institutions of higher education participated in the design, development, piloting, and field testing of edTPA from 2009 to 2013. edTPA has been used operationally to assess teacher candidates since Fall 2013 and is now used by nearly 700 programs in 38 states. **edTPA is the first subject-specific, standards-based pre-service assessment and support system to be nationally available in the United States.**

Role of the Partners

edTPA was created with input from teachers and teacher educators across the country in a process led by Stanford University's Center for Assessment, Learning and Equity (SCALE) and supported by AACTE.

Each of the edTPA partners supports edTPA development and implementation in different ways. Stanford University faculty and staff at SCALE developed edTPA and are the sole authors. They receive substantive advice and feedback from teachers and teacher educators. The national design team and individual subject-specific design teams were convened annually to develop and update the handbooks for each of the 27 teaching fields. Design team members included subject-matter organization representatives from higher education and P-12.

As the lead in development, 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 (including benchmarks), 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 (see description below).

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 public 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.

The design framework for edTPA and constructs assessed were established prior to the partnership with Evaluation Systems/Pearson and were informed by earlier experiences and work led by SCALE staff (National Board and PACT). Evaluation Systems was chosen as the operational partner to ensure that edTPA assessment development built by the profession and supported by foundation funds could be scaled up for national use. That is, the Evaluation Systems/Pearson group has no authority or decision-making role in the design and development of edTPA.

edTPA as Support and Assessment

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 (Pecheone & Whittaker, 2016).

Summary of resources

Since edTPA launched its first "online community" in 2011, membership has grown to 9,082 faculty from approximately 700 teacher preparation programs who have access to more than 165 resources including candidate handbooks, rubrics, and 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 676,000 times.**

The most commonly downloaded resources include:

Understanding Rubric Learning Progressions - Full Collection	39996 downloads
edTPA Handouts to Share with Stakeholders	32962 downloads
2013 edTPA Field Test: Summary Report	15556 downloads
2014 edTPA Administrative Report	13999 downloads
State Policies & EPP Implementation Support Chart	12688 downloads
edTPA Guidance To Supervising Teachers	12412 downloads
Elementary Education Handbook	12151 downloads
Special Education Handbook	11568 downloads
edTPA Myth Busters	11295 downloads
Secondary Science Handbook	11122 downloads

In addition to the Resource Library for edTPA members, the website also includes an online community platform used by faculty to pose questions or share resources developed locally.

Since the 2014 Administrative Report was published, SCALE has revised and updates numerous resources that can be used by candidates and programs to not only provide formative opportunities to develop and practice the features of effective teaching assessed by edTPA, but also to help candidates interpret their edTPA performance and for faculty to understand edTPA results as actionable evidence. Several examples are described here:

- **Making Good Choices** - For candidates to review as they prepare their portfolios; support guide for navigating edTPA and preparing artifacts and commentaries for submission

- **Local Evaluation rubrics & samples** - Available to programs/coordinators/faculty who have either completed the online orientation to local evaluation or attended an in-depth local evaluation workshop. The materials are available through secure online system and can be used to engage faculty and P-12 partners in examining and providing feedback to candidates.
- **Understanding Rubric Level Progressions (URLP)** - For programs to use when guiding candidates towards a deeper understanding of what evidence of beginning teacher practice looks like at each rubric level, building across each rubric progression
    - Available in 27 versions representing all edTPA handbooks, the resource is highly similar to the Thinking Behind the Rubrics used by scorers who evaluate edTPA portfolios, the URLP is designed to help faculty and candidates understand edTPA expectations.
- **Guidance for Acceptable Support** - Provides examples of acceptable vs. unacceptable support for programs to provide to candidates. A supplemental resource is also available on the "Educative Use of edTPA Materials" and how they can be examined for formative purposes and faculty/peer feedback.
- **Review of Low Scoring edTPAs** - Describes common reasons for low performance and how to address them. Provides guidance to inform retakes.
- **Retake Guidelines** –Describes how to interpret candidate evidence and support candidates in retaking the edTPA -- includes instructions for resubmission for each task as appropriate
- **Professional Growth Plan** –A tool for helping candidates integrate edTPA, along with other sources of data, into their professional development plan as a beginning teacher.

National Academy

edTPA's National Academy of consultants provides onsite professional development and implementation support for programs, states, and regional networks, as well as webinar-based support for individual programs seeking more peer interaction. National Academy members must demonstrate edTPA leadership within a program, have experience leading state or local implementation and/or developing and delivering edTPA-related professional development, and have disciplinary expertise related to national scoring and training. Since the National Academy launched in early 2015, nearly 100 workshops and events have been supported.

Common workshop topics include:

- General introduction to edTPA
- "Deep-dive" handbook and rubric walk-throughs
- Preparation for local evaluation
- Curriculum inquiry
- Academic language
- P-12 support
- Candidate support
- Leading faculty in a change process

SCALE collects feedback from each workshop to inform continual improvement of the National Academy, which is intended to be an adaptive and responsive resource addressing programs' evolving needs.

Semi-Annual Summary Reports

edTPA Summary Reports are made available to Educator Preparation Programs (EPPs) and state agencies on a biannual basis (January and July) to assist them in examining the performance of their candidates as compared to the population of candidates taking edTPA within the associated state and nationally. The reports provide analyses at three levels for the date ranges referenced:

edTPA National Performance Summary

Provides a summary that represents national-level data for candidates scored and reported within the stated date ranges. Programs who have received edTPA official data in these date ranges will receive this summary.

edTPA State Performance Summary

Provides a summary that represents state-level summary data for candidates who indicated they were prepared in the state, and were scored and reported within the stated date ranges. Programs who have received edTPA official data in these date ranges will receive this summary for their respective state.

edTPA EPP Performance Summary

Provides a summary that represents program-level summary data for candidates who indicated they were prepared at the specific program, and were scored and reported within the stated date ranges. Programs who have received edTPA official data in these date ranges for candidates preparing at the program will receive this summary for their program.

All summary reports contain: a) mean edTPA scores, total and by rubric, b) distributions of total scores, and c) rubric means and distributions for each field. In addition to the three summary reports, EPPs are provided a spreadsheet or roster that provides official scores by rubric as well as total scores by task and overall for each candidate who indicated they were prepared by the program and was officially scored and reported during the stated date ranges. The report allows the EPP to easily analyze performance by subject area, cohort, or other program features.

EPPs utilizing the data are also provided with a detailed table of contents and suggested questions to guide conversation about each part of the reported data. Examples of questions include: "What do the data show in terms of teacher candidates' understandings and professional performance? What are the implications for our program in terms of what and how we teach?" SCALE encourages programs utilizing the data to connect numerical trends to local evaluation of candidate portfolios.

edTPA National Condition Code Report

A National Condition Code Report is available annually to provide programs with more information to inform curriculum and knowledge of current trends in edTPA submissions. The report summarizes the condition codes that were applied during the operational year, providing the frequency of condition codes by field and by condition code reason. In order for a candidate's edTPA submission to be scored, it must meet assessment Submission Requirements. If a submission does not meet the requirements and the submission or portions of the submission are unable to be evaluated by a scorer, a "Condition Code" will be assigned indicating the requirement(s) that have not been met for that particular rubric. Condition codes are assigned when materials do not meet the submission guidelines (e.g., wrong file format, file is unreadable, video has been edited). The Submission Requirements provide examples of reasons why a particular condition code may be assigned. The most recent Condition Code Report identified that the total number of portfolios assigned one or more condition codes represents less than 5% of total edTPA submissions.

These reports are critical to building understanding and discussion about edTPA, and for this reason, SCALE strongly encourages EPPs to share these data with all participating faculty and P-12 partners to celebrate candidate success and as part of ongoing program renewal conversations.

Evaluation Systems/Pearson Supports

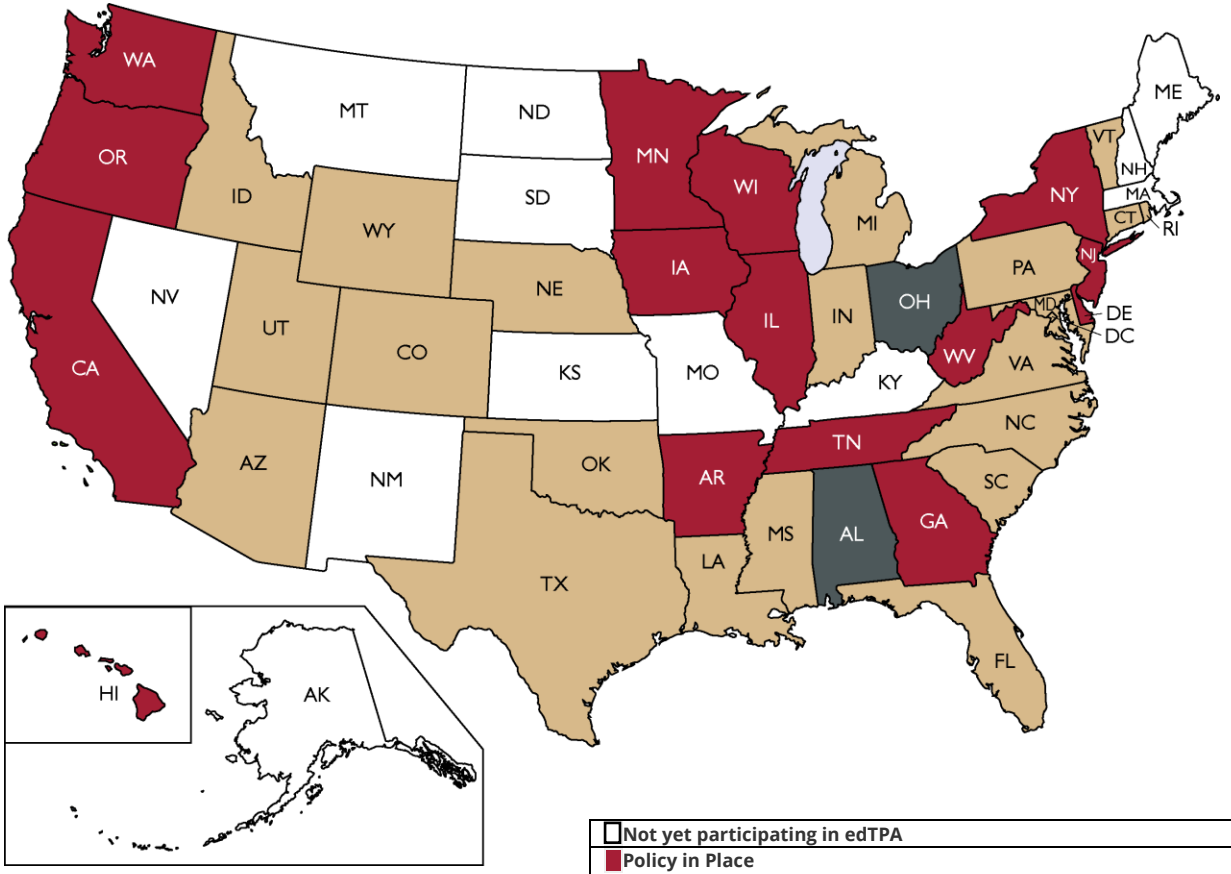
Pearson (through edTPA.com – the candidate-facing program web site) provides operational assessment services associated with registration, scoring, and reporting of edTPA scores. Assessment services include use of the technology platform which registers the candidate, receives the portfolio, coordinates the logistics of scoring the portfolio, and reports the results to the candidate. Additionally, a faculty feedback feature is available through the Pearson Portfolio system, allowing candidates to request formative feedback from a designated faculty member based on SCALE's guidelines of acceptable support. Assessment services also include the recruiting and management of qualified educators who serve as scorers, scoring

supervisors, or trainers. Scorers are trained using a training curriculum developed by SCALE, specifically for use with edTPA rubrics. Scorers use standardized scoring procedures and are calibrated and monitored during scoring. Pearson also works with EPPs and state agencies to securely report candidate scores as appropriate. Through the *ResultsAnalyzer* tool, stakeholders are able to review and utilize their data sets as provided on each reporting date.

Pearson also provides fee waivers in the form of financial hardship vouchers to eligible candidates. Over 3,105 fee waivers were made available for eligible edTPA candidates between September 2013 and June 2016. Waivers are provided directly to State Agencies and/or EPPs who then distribute them based on student need.

States Participating in edTPA

The map below shows the states currently participating in edTPA as of August 2016, signaling those with an edTPA implementation policy in place and those expecting to have an implementation policy in place soon. Visit edTPA online at http://www.edtpa.aacte.org/ for up-to-date information on states and their policies.



Not yet participating in edTPA

Policy in Place
In general, these states have statewide policies in place requiring a state-approved performance assessment as part of program completion or for state licensure and/or state program accreditation/review. In these states, edTPA also has been approved as a performance assessment for these purposes.

Taking Steps Toward Implementation
A performance assessment and/or edTPA are been considered at the state level for program completion or as a licensure requirement.

State Participating in edTPA
At least one provider of teacher preparation-either traditional or alternative-is exploring or trying out edTPA.

edTPA Scoring 2015

Over 2,500 teachers and teacher educators now serve as trainers, scoring supervisors, or scorers of edTPA as part of the National Scoring Pool. Scorers must be P-12 teachers or teacher preparation faculty (Including adjuncts and clinical supervisors) with significant pedagogical content knowledge in the field in which they score, as well as experience working as instructors or mentors for novice teachers. In the 2015 administrative year (January 1st, 2015 – December 31st, 2015), recruitment goals targeted a balance of scorers with approximately 50% teacher educators and 50% classroom teachers. Of these qualified scorers, 32% of the practicing classroom teachers and 20% of the qualified scoring pool are National Board certified teachers.

Scorer Training

Before becoming an official edTPA scorer, educators must go through an extensive scorer training curriculum and meet qualification standards. All scorer training materials are authored or reviewed by SCALE. Training for scorers comprises both individual online and interactive group sessions, totaling about 20 hours. The individualized training includes a series of modules that orient scorers to the tasks, rubrics, and scoring system, and provides numerous opportunities to identify and evaluate evidence for each rubric. After completing the individual portion of the training materials, scorers independently score a sample edTPA portfolio coded by experienced scorers and trainers and then review evidence and score justifications with other scorers and a trainer in that content area. Following the independent sample scoring of a practice portfolio and explanations for score justifications, scorers must consistently score two qualifying portfolios within calibration standards before becoming fully qualified to score.

Low-Incidence Fields:

The following fields have low candidate volumes and follow a slightly modified training plan that includes several online modules and webinar based meetings with trainers and other scorers to discuss several practice portfolios that have been consensus scored:

1. Agriculture Education
2. Business Education
3. Educational Technology Specialist*
4. Family and Consumer Science
5. Health Education
6. Library Specialist
7. Literacy Specialist*
8. Technology and Engineering Education
9. Classical Languages*

The three fields marked with an asterisk are consensus scored due to their very low number of portfolio submissions. Consensus scoring consists of a process whereby two or three scorers meet with a trainer/facilitator to score the same portfolio and to arrive at a consensus score for each rubric. Following the training, fully qualified, active scorers in all fields are monitored by their supervisors through a back-reading process and routinely score previously scored "benchmark" portfolios to ensure they are applying scores accurately and consistently.

Scorers are recruited, trained, and qualified to score in two scoring pools – national and regional (see additional information in the "Regional Scoring Option" section below). The national pool includes qualified scorers who access and score portfolios submitted from across the country. In the regional scoring pool, qualified faculty from preparation programs (in implementing states where regional scoring is an accepted scoring model), score a sample of their program's own candidate portfolios. Regional scorers complete the same training and qualify using the same criteria before scoring, and have the same quality monitoring and scoring consistency requirements as those scoring in the national pool. Additionally, portfolios scored by regional scorers are double scored by the national pool.

Each edTPA scorer is assigned to score portfolios at the grade-level span and subject area for which he or she has qualified. The scorer utilizes a secure

online scoring platform to access each candidate's materials and determines the rubric scores after viewing all evidence from artifacts, commentaries, and video recording(s) submitted by the candidate. Drawing upon SCALE's theory of action from PACT that examined the benefits of understanding the interrelationships within a cycle of effective teaching, each scorer scores an entire candidate submission (rather than independent scorers of discrete tasks or rubrics). As a result, the scorer can effectively review the entirety of a candidate's teaching evidence and ensure the components are appropriately interrelated. The scorer evaluates how the candidate **plans** to support subject-specific student learning, **enacts** those plans in ways that develop student learning, and **analyzes** the impact of that teaching on student learning.

Guided by 15 analytic rubrics (five rubrics within each of the three assessment tasks) that use a five-point scale, the scorer assesses the extent to which — and the areas in which — the candidate is ready to teach, as well as any particular areas for improvement. The total possible scores on edTPA for fields with 15 rubrics, added across all 15 rubrics, range from 15 to 75 points.

edTPA's Scoring Model

Overview of the edTPA Scoring Model:

- Scorers evaluate the entire portfolio.
- Rubric scores are on a five-point scale – rater agreement is evaluated by exact and adjacent scores.
- Scoring model: currently about 30% of portfolios are double scored, for two reasons:
    1. 10% of portfolios are randomly selected for reliability reads OR
    2. The portfolio lies within the double scoring band around the state or national cut score.
- Inter-rater reliability is calculated by examining the double scored portfolios cited under #1 above (10% reliability reads).
- If a portfolio score falls within the double scoring band (a band calculated based on the standard error of measurement around a state cut score or the national recommended professional performance standard), it is scored by a second scorer.
- Double scored portfolios can be read by a scoring supervisor (a third "chief" scorer) for rubric score resolution, or for portfolio score adjudication.
    - Resolution: If Scorer 1 and Scorer 2 are discrepant (i.e., more than 1 score point apart) on any rubric, the portfolio is resolved by a scoring supervisor. The supervisor score is reported for the discrepant rubrics.
    - Adjudication: If Scorer 1 and Scorer 2 are on opposite sides of the national recommended professional performance standard, the portfolio is adjudicated by a scoring supervisor who scores the entire portfolio. The scoring supervisor scores are reported to candidates.
- If a portfolio is double scored and does not need resolution or adjudication, then the average of scorer 1 and scorer 2 is reported to the candidate.

The double scoring procedures increase the decision consistency of the final scores assigned to edTPA candidates. In all such cases the final score is based on at least two scorers who agree on the decision in relation to the state cut score or the national recommended professional performance standard. Ideally, decisions of the two scorers on each of the 15 rubrics would be the same across the portfolio. However in practice, the high complexity of teaching and 15 different decisions by rubric may result in a difference in total scores across two raters. Evidence of high total agreement (the rate at which scorers assign the same or adjacent scores) presented in the '*Reliability*' section of this report supports the consistency of edTPA scores.

Scoring for edTPA occurs year-round, with results typically reported approximately every two weeks. Given this ongoing scoring model (as contrasted with a single, event-based scoring session), scorer quality monitoring is in place on a constant basis. Facets of the quality management of scorers include:

- **Validity Portfolio Performance:** Validity portfolios are benchmarked portfolios (i.e., calibration exercises) that are randomly sent to scorers to evaluate scorer performance. Approximately 10% of the portfolios a scorer sees are validity portfolios.
- **Inter-Rater Reliability:** As described above, 10% of portfolios are randomly double-scored to monitor agreement rates amongst scorers.
- **Monitoring after Initial Qualification:** All newly qualified scorers are backread by a scoring supervisor. All scorers are flagged for backreading after they have scored their first portfolio.
- **Scoring Rate:** Scorers are monitored to ensure they are not scoring too quickly or too slowly, which may impact quality. On average, a portfolio is scored in 2-3 hours. A scorer's average scoring rate per portfolio cannot not exceed or fall below edTPA program thresholds.
- **Excessive Scoring:** Scorers are not permitted to score an excessive number of portfolios in a designated time period.
- **Portfolio Limits:** The edTPA program limits the number of portfolios in each subject area that any individual scorer may score during a specific timeframe.
- **Backreading:** Scorers are systematically monitored by their supervisors through a backreading process that ensures they are applying scores accurately and consistently. Backreading is defined as supervisors scoring a previously scored portfolio for the purpose of reviewing the original scoring and providing feedback to the scorer. During backreading, a scoring supervisor applies scores and identifies key evidence to support the scores. After applying scores, supervisors review scores from the original scoring and review backreading scores with feedback to the original scorer.
- **Period of Inactivity:** Inactive scorers (those who have not scored within 120 days) need to score a complete benchmarked portfolio as a re-qualification exercise in order to remain calibrated to edTPA rubrics and prior to returning to score.

Regional Scoring Option

Faculty engagement in the scoring of edTPA portfolios is an ideal way to deepen and sustain an understanding of candidate performance and educative implementation. In addition to faculty participation as scorers in the national official scoring outlined above, EPPs can participate in regional official scoring, wherein faculty are able to officially score portfolios from their own campus or region.

Regional scorers complete the same training and qualify using the same criteria as all official scorers before scoring, and have the same quality monitoring and scoring consistency requirements as those scoring in the national pool and as described above. edTPA regional scoring is conducted in accordance with all quality standards in place for national scoring, to ensure that the levels of service and quality of the national program are maintained. These quality standards refer to both the actual scoring statistics and figures, as well as scorer training quality protocol. Scorers observe all conditions and requirements for training and qualification, as well as of confidentiality and self-recusal for personal knowledge of the candidate.

The regional scoring option was piloted in spring 2015 in California and was made available to all edTPA Implementation Members in a second comprehensive pilot phase in spring 2016, in order to establish processes for a broad-based implementation of edTPA regional scoring. Further piloting will be done in Spring 2017, and based on the results of the pilot, a complete national expansion will be offered in 2018 (scoring occurring in Spring 2018).

The EPP will play a primary role in the management and implementation of regional scoring on their campus. The number of faculty from the EPP who complete scorer training and qualify will determine the number of portfolios that can be identified for regional scoring at the location during specified scoring windows.

It is hoped that regional scoring will offer EPPs additional opportunities to build faculty capacity to support prospective teachers as well as become

more engaged and knowledgeable about edTPA handbooks, the scoring process, and performance of candidates.

Candidate Submissions, Originality, Score Confirmation, and Retakes

At the time of the submission, edTPA candidates are required to attest to the originality of their work, including confirmation that the candidate is sole author of the commentaries and other written responses to prompts and other requests for original information in this assessment, and that the candidate has appropriately cited all materials in the assessment whose sources are from published text, the Internet, or other educators. Pearson uses a well-established and reliable software platform to screen submissions for originality of content. Submissions that are flagged as a result of initial screening are subject to additional review and investigation in coordination with individual IHEs or state or, as appropriate. In some cases, the release of a candidate's edTPA results may be delayed as the result of an administrative review. An administrative review may occur for several reasons, including confirmation that a submission meets all requirements and is in compliance with the rules of assessment participation.

Following score reporting, if a candidate believes that one or more of their scores has been reported in error, they may request a score confirmation. The score confirmation process involves having a supervisor or trainer who did not serve as one of the original scorers, review the original reported scores to confirm that they are accurate. As the supervisor or trainer conducts their review, should there be a score with which the supervisor or trainer disagrees, they rescore the entire portfolio and provide the updated rubric scores.

If the score confirmation process results in a score alteration, the candidate is issued an updated Score Profile, the score confirmation fee is refunded, and the candidate's records will be updated. If the original score is confirmed as a result of the score confirmation process, the candidate is sent a letter indicating that their score has been confirmed, and the score confirmation fee is not refunded.

Candidates who do not meet their educator preparation program or state requirement may retake the assessment by choosing from either retaking the full assessment, or by retaking single or multiple tasks. The edTPA Retake Instructions for Candidates provide important information on the process of retaking and materials necessary for a retake submission.

Validity Evidence

According to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014) and leading psychometric experts (Bell et al., 2012; Haertel, 2008; Haertel & Lorié, 2004; Kane, 2006; Sheppard, 1993), the process of validation begins with defining the intended purpose of the assessment and the constructs being measured. The inferences made by this definition are then examined using various sources of validity evidence that may support the interpretation and use of scores. edTPA was developed to be an authentic, subject-specific, performance-based support and assessment system of a candidate's initial readiness to teach. The following section of the report presents the inferences made by this purpose and use of edTPA, followed by evidence that evaluates the validity of proposed score interpretations.

Content Validity and Job Analysis

edTPA was designed following standards for credentialing exams, and intended to be used as an assessment of the knowledge, skills, and abilities necessary for beginning teaching. According to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), "validation of credentialing tests depends mainly on content-related evidence, often in the form of judgments that the test adequately represents the content domain associated with the occupation or specialty being considered." The AERA, APA & NCME Standards (2014) indicate that, "To identify the knowledge and skills necessary for competent practice....A wide variety of empirical approaches may be used, including the critical incident technique, job analysis, training

needs assessments, or practice studies and surveys of practicing professionals." Building on the foundation of NBPTS, PACT, and InTASC, the development of the edTPA rubrics was informed by a combination of content validation and job analysis activities and information. The information obtained through these activities is a key contributor to validating edTPA as an effective, authentic instrument that can be used for teacher licensure decisions. The review by teachers and teacher educators provided statistical data to support edTPA as a highly representative tool in measuring candidates' knowledge and skills needed to perform on the job as a novice teacher. The data support edTPA as an evaluation tool for both pedagogical and subject-specific knowledge and skills — which together with other measures of teacher competence form the basis of what teacher candidates must possess starting on day one of their professional career.

To further support the content validity findings in 2013, a confirmatory job analysis study was conducted to support the job-related validity of edTPA by drawing upon the list of Knowledge, Skills, and Abilities (KSAs) that were identified by educators, faculty, and subject-matter experts during the edTPA development process. Subject-matter experts for edTPA, composed of teachers and/or educators who train those entering the profession, generated the following list of KSAs:

1. Planning for content understanding
2. Planning to support varied student needs
3. Planning assessments to monitor and support student learning
4. Demonstrating a positive and engaging learning environment
5. Engaging students in learning
6. Deepening student learning while teaching
7. Subject-specific pedagogy
8. Analyzing student work
9. Providing feedback to guide learning
10. Supporting students' use of feedback
11. Using knowledge of students to inform planning
12. Analyzing teaching
13. Using assessments to inform instruction
14. Identifying and supporting language demands
15. Using evidence of language use to support content understanding

These edTPA KSAs served to inform refinements to the design and development of edTPA. The assessment instruments' tasks and scoring rubrics directly align to these KSAs. As a form of confirmatory evidence, job analysis activities were conducted to examine the links between these KSAs and teachers' actual work. The job analysis confirmation serves as evidence supporting the validity of the interpretations made based on the edTPA results.

Through this process the 15 core edTPA rubrics were confirmed as representing knowledge, skills, and abilities that are judged to be important or critically important to perform the job of a teacher as represented on the job related survey.

For a full overview of the Content Validity and Job Analysis evidence gathered in edTPA development, please refer to the 2014 Administrative Report.

Construct Validity

Based on this foundation and design process, edTPA is a subject-specific performance assessment that evaluates a common set of teaching principles, teaching behaviors, and pedagogical strategies. The rubrics of the assessment are divided into three tasks that assess the integrated cycle of planning, instruction, and assessment that underlies teaching. Exploratory Factor Analyses (EFA) of 2013 field test data provided support for the common underlying structure of edTPA that unifies all rubrics, as well as for the three-task structure (see pg. 22 of the 2013 edTPA Field Test Summary Report). Confirmatory Factor Analyses (CFA) as well as a Partial Credit IRT model were conducted using data from portfolios submitted in 2015, both described in the "Internal Structure" section below. Both of these models confirmed that the tasks are measuring a common unifying teaching construct and that there are three common latent constructs (planning, instruction, and assessment) that are appropriately assessed by the rubrics

which make up each of the three tasks. These analyses confirm the intended design and structure of edTPA and provide evidence that edTPA scores measure key job-related teaching skills that are used to evaluate a candidate's overall readiness to enter the profession of teaching.

In addition to the evidence presented in the Field Test Summary Report and described above, the edTPA Review of the Research, developed by SCALE staff with input from educators and researchers, is a resource that identifies foundational research literature that informed the development of edTPA and ongoing validity research. The extensive literature review cited provides a foundation for the common edTPA architecture used across 27 different subject-specific licensure/certification areas and the fifteen shared rubric constructs that define effective teaching. The document includes foundational texts in the field relevant to each performance task (planning, instruction, and assessment) and rubrics. The studies cited provide an empirical examination of the constructs including reviews that summarize the state of the research evidence in that field, and professional papers, chapters, and books that make research-based recommendations for practice. The first section of the review presents relevant literature and research that speaks to the role of assessment in teacher education and student learning. The sections following are organized according to the three edTPA tasks (planning, instruction, and assessment), and by rubric within each task and provide a strong basis for the teaching competencies used in edTPA.

Consequential Validity

edTPA is intended to be embedded in a teacher preparation program as an educative tool and support system for candidates, faculty, and programs. Evidence of validity, then, must come from examining how use and implementation of edTPA impact program curricula, faculty, and teacher candidates.

Numerous scholars have outlined the benefits of high-quality formative performance assessment and the opportunities for improvement that common standards, experience of implementation, and use of data gathered can provide (e.g., Darling-Hammond, 2010; Darling-Hammond & Falk, 2013; Pecheone & Chung, 2006; Peck, Gallucci, Sloan, & Lippincott, 2009; Peck, Singer-Gabella, Sloan, & Lin, 2010; Sato, 2014).

Several studies have now verified these claims using their experience with edTPA as well as PACT, the precursor to edTPA that shares the same architecture and assesses many of the same constructs. Reports by these programs indicate that thoughtful integration of PACT/edTPA knowledge, skills, and constructs into pre-service preparation programs has improved the content, methods, and supports of program curriculum (Gillham & Gallagher, 2015; Peck & McDonald, 2013; Sloan, 2013). The use of PACT and edTPA has been reported to support program improvement and inquiry, collaboration within and between institutions around program structure, practice, and quality, as well as reflection on teacher candidates' performance and needs (Chung, 2008; Kleyn, Lopez, & Makar, 2015; Liu & Milman, 2013; Peck, Gallucci, & Sloan, 2010; Sloan, 2013; Stillman, Anderson, Arellano, Lindquist Wong, Berta-Avila, Alfaro, & Struthers, 2013).

edTPA enables programs to clearly communicate expectations to students, and to engage in conversations and collaborations across programs and institutions using a common language. These studies also report some challenges or unintended consequences experienced by programs, faculty, and candidates as they work to integrate edTPA requirements into existing practice and navigate the pressures that come with high-stakes policy – findings that are well documented in student assessment. However, edTPA was designed as a support and an assessment program and targeted attention to capacity building and implementation was explicitly built into the system to help mitigate the high-stakes use of edTPA — from a system of compliance to a system of inquiry.

Policy and approach to implementation play important roles in the impact of the assessment on the program and the teacher candidates' experiences (Peck, Gallucci, & Sloan, 2010; Whittaker & Nelson, 2013). A recent study has found that candidate engagement with these opportunities to learn implicit in the process of taking edTPA are mediated by the attitudes and actions of faculty, cooperating teachers, and field supervisors (Lin, 2015). Evidence

supports the inference that despite challenges and workload, teacher candidates report that constructing their PACT/edTPA portfolios has expanded their understanding of pedagogy and assessment of student learning, caused them to reflect more deeply on their instruction, and that they expected this experience to be useful to their future practice (Chung, 2008; Darling-Hammond, Newton, & Chung Wei, 2013; Lin, 2015).

Concurrent Validity

Evidence of concurrent validity examines the inference that edTPA scores accurately reflect a candidate's readiness to teach by testing whether total scores are related to other indicators of instructional capability. Empirical examinations of this type of evidence require datasets with a substantial sample size that include variables from various measures of performance, as well as variables that allow for the control of other sources of variance such as demographic categories and prior skills and knowledge. These studies are now beginning to emerge: a study from Illinois State University has found that candidates' edTPA scores correlate with GPA, scores on a content knowledge assessment, and scores on a pedagogy and skills assessment (Adkins, Klass, & Palmer, 2015). Findings presented later in this report also indicate that demographic variables are not associated with differences in edTPA scores. Another study that focused on supervisors' predictions about their candidates' performance on PACT found that these predictions accurately predicted PACT scores (Pecheone & Chung, 2006). As programs gather more data, several studies around the country are being conducted that will add to this collection of evidence. SCALE is currently working on a state-wide concurrent validity study with the state of Georgia to examine the relationship between edTPA scores and other markers of performance completed during pre-service teacher preparation that can provide evidence of convergent and divergent validity, as well as interactions with demographics, program type, and degree type. Dissemination of these results as they become available will inform all programs and states working with teacher candidates taking edTPA.

Predictive Validity

Licensure assessment is designed to assess core skills and abilities in teaching and learning that are aligned to professional standards, research, professional practice, job related skills and wisdom of practice. Predictive Validity studies (routinely conducted after the assessment has been in operational use for several years) provide another method of validating the use of edTPA scores as markers of readiness to teach by examining their ability to predict student learning and instructional practice on the job, however we must exercise caution in not narrowing and marginalizing effective teaching. While valuable, predictive validity studies do not address the relationships of preparation with other known measures of teacher effectiveness (teacher evaluation, impact of mentoring, impact of culturally relevant pedagogy etc.). Finally, licensure testing is a threshold measure (i.e., a demonstration of a minimum competency to be ready to teach), as contrasted with a highly effective teacher that could impact student learning – which is a demonstration of a much higher bar than entry level cutoff scores. SCALE does not oppose conducting predictive validity studies as part of a comprehensive study of teaching, but the data need to be interpreted with great caution and with respect to the VAT research literature and the low effect sizes. SCALE supports the use of predictive validity studies as one part of a comprehensive construct validity study.

Predictive validity evidence for PACT was revealed in a study by Darling-Hammond, Newton, & Chung Wei (2013), which found that teachers' PACT scores predict growth in their students' math and literacy achievement using value-added statistical modeling. Preliminary data from studies by Benner and Wishart (2015) has revealed that edTPA scores predict candidates' ratings of teacher effectiveness, as measured by a composite score that combines students' performance data and classroom observations. More recent data reported at the May and August 2016 meetings of the Tennessee Board of Education subcommittee on educator preparation and licensing demonstrated that candidates with higher scores on edTPA were also more likely to have higher ratings on the TN teacher evaluation system which includes supervisor observation evidence and student learning measures.

Further, a recent study by Goldhaber, Cowan and Thoebald (2016) used teacher candidates' scores on edTPA (from the field test and first operational year) to provide estimates of the extent to which edTPA performance is predictive of the likelihood of employment in the teacher workforce and value-added measures of teacher effectiveness. They found that edTPA scores were "highly predictive of employment in the state's public teaching workforce, and evidence on the relationship between edTPA scores and teaching effectiveness was more mixed. Specifically, continuous edTPA scores are a significant predictor of student mathematics achievement, but when edTPA was a binary screen of teaching effectiveness (i.e., pass/fail), passing edTPA was significantly predictive of teacher effectiveness in reading but not in mathematics." These results are consistent with VAM studies conducted on the National Board and PACT.

In addition, the Education Policy Initiative at Carolina (EPIC), in partnership with the UNC General Administration and the 15 UNC system institutions engaged in teacher preparation, has established and is continuing a body of research to assess the construct validity, reliability, and predictive validity of both locally and officially-evaluated edTPA portfolios. This work initiated with analyses of locally-evaluated TPA portfolios from the 2011-12 graduating cohort at one UNC system institution. These results are available as a working paper on the EPIC website (http://publicpolicy.unc.edu/epic-home/) and published in *Teaching and Teacher Education.*

In fall 2016, EPIC produced a policy brief (https://publicpolicy.unc.edu/files/2016/10/Initial-Findings-from-edTPA-Implementation.pdf) summarizing edTPA implementation in North Carolina, detailing how UNC system candidates are scoring on edTPA, and assessing the construct validity and predictive validity of officially-evaluated portfolios. These predictive validity analyses focus on the 2013-14 graduating cohort of one UNC system institution who went on to be first-year teachers in the 2014-15 school year. Importantly, these predictive validity analyses focus on first-year teachers' value-added estimates and evaluation ratings. Overall, these predictive validity results show that edTPA measures significantly predict first-year teacher performance. Concerning teacher value-added, 7 of 15 edTPA rubrics are significantly associated with a standardized measure of teacher effectiveness; summatively, the standardized edTPA total score and having a total score of 42 or greater also predict significantly higher value-added estimates. Regarding teacher evaluation ratings, the edTPA Instruction construct predicts significantly higher evaluation ratings on 4 of 5 teaching standards; the Assessment construct predicts significantly higher evaluation ratings on 2 of 5 teaching standards. At the edTPA rubric level, many rubrics, particularly in the Instruction construct, predict significantly higher evaluation ratings. Lastly, the two summative edTPA measures—the standardized total score and scoring at 42 or greater—predict significantly higher evaluation ratings for 3 of 5 teaching standards. More data are needed—from additional universities and graduating cohorts—to replicate these results.

Likewise, in fall 2016, EPIC will release a working paper that illustrates a two-pronged empirical framework—latent class analysis and predictive validity analyses—that teacher preparation programs can use to analyze their edTPA data for program improvement purposes. With new consequential policy for edTPA and expanding use in North Carolina—several universities have edTPA scores beginning with their 2014-15 graduating cohort—EPIC will continue analyses throughout the 2016-17 academic year. These analyses, expected in the spring/summer of 2017, will assess the predictive validity of officially-evaluated edTPA portfolios from multiple UNC system institutions.

Regarding teacher evaluation ratings, the edTPA Instruction construct predicts significantly higher evaluation ratings on 4 of 5 teaching standards; the Assessment construct predicts significantly higher evaluation ratings on 2 of 5 teaching standards.

As mentioned above, predictive validity studies are not a precursor to implementation of licensure assessments of teacher candidates, as it is not possible to analyze predictive validity during clinical practice, as candidates are not the teacher of record during this time. Additionally, analyzing these relationships requires gathering data on a sample that is large enough to determine consistent, generalizable patterns (as with the UNC and Goldhaber studies). Once candidates become teachers of record, the examination of predictive validity is more robust if researchers are able to follow candidates into their teaching practice for several years in order to obtain more stable estimates of student learning and teacher effectiveness as captured by student test scores and other assessments of performance, (e.g., observations of teaching practice, classroom climate surveys, supervisor, co-teacher, student, peer evaluations). SCALE, and state level partners like those in Georgia and North Carolina, are committed to conducting predictive validity studies that follow candidates into employment if the state database enables linking teachers to classrooms and student achievement – providing states grant access to these data. The edTPA National Technical Advisory Committee of leading psychometricians in the field advises SCALE on the design of studies that examine the impact of edTPA implementation as an assessment and educational tool on educator preparation programs, faculty, candidates, P-12 educators, and P-12 students' achievement. The newly convened edTPA research group comprised of faculty representatives across states using edTPA work with SCALE to identify and collaborate on research efforts relevant to teacher education.

Internal Structure

The use of edTPA rubric, task, 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 weaknesses of a particular candidate or a preparation program. Scores on each of the rubrics 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 (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 valid 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 following analyses of the internal structure of edTPA provide psychometric evidence that support 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 the "edTPA Scoring 2015" section above.

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). These analyses were provided in the 2014 Administration Report, and to confirm these factor structures again, Confirmatory Factor Analyses (CFAs) were conducted using data from operational portfolios submitted in 2015. CFAs 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 27,759 first-time edTPA submissions, and excluded incomplete portfolios and portfolios with condition codes.[3] In cases where a portfolio was double-scored, only the first rater's score is included in the analyses. CFA models were estimated based on the observed sample covariance matrix among rubric scores for the 2015 administration cycle. Models were estimated using maximum likelihood estimation with standard errors and scaled chi-square fit statistics, as implemented in the R package "lavaan" (Rosseel, 2012), to fit all models.

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 was also estimated. All factor loadings for the three-factor solution were positive and statistically significant as anticipated (all standardized loadings were greater than .6 in the 3-factor model). All but one of the factor loadings for the one-factor solution attained magnitudes of at least 0.50, with just a single rubric (Rubric 6) with a factor loading slightly below that target (0.496). 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 measure three highly interrelated sub-dimensions – planning, instruction, and assessment – of a single readiness to teach construct.

IRT: Partial Credit Model

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 PCM been used to evaluate the internal structure of similar portfolio-based assessments of readiness to teach such as PACT (Duckor et al., 2014). In the PCM the underlying theta variable is a direct function of the total score, which allows the theta score to function as a statistical representation or summary of "readiness to teach" as measured by the total sum score on edTPA. The PCM thus provides information about the relationship between candidates' readiness to teach as measured by a total sum score and edTPA rubrics consistent with the edTPA policy for summing across rubrics and subject area fields to evaluate candidate performance.

It is important to note that this model was used to further examine the theoretical foundation that underlies the use of edTPA total scores as a representation of a common construct of teaching effectiveness, and that the rubric levels are distributed in the expected pattern of difficulty. edTPA scores are not derived using IRT analyses; total scores are an aggregate of all rubric scores across the assessment. The dataset analyzed here contains a single score for each candidate and this single score is derived from the ratings of a single scorer. edTPA rubrics were designed to be independent measures of the teaching constructs measured in edTPA; it is possible that the rubric scores may be affected by the presence of some individual rater effects due to the single scorer approach used to score edTPA. However, the design of edTPA is a reflection of a theory of action that is grounded in the licensure approach and over a decade of experience with the InTASC

[3] 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.

portfolio and PACT program in California which is designed to provide higher education faculty with a comprehensive profile of a candidate's performance within an authentic and interconnected cycle of teaching.

Finally, the results presented below are based upon aggregating data across credential areas. Again, edTPA is used based on a single total score calculated equally across fields and so this analysis provides evidence about how this measure functions overall. However, we also plan to explore in future analyses l fit models separately by credential areas. We note, however, that there are not enough candidate submissions in most edTPA credential areas to fit the PCM with stable estimates. A primary limitation is that as sample sizes become smaller, there are sometimes no observed scores in all possible categories for all rubrics, and not all relevant parameters can be estimated. As more candidates complete edTPA, further analyses by subgroups will become more possible.

The PCM was used to investigate the following primary questions:

- How well does a unidimensional PCM fit edTPA data?
- Do all rubrics adequately fit the model?
- Are the rubric score-point thresholds distributed across the latent theta distribution, suggesting the rubrics are well-matched to the candidate performance distribution and provide a good measurement of each candidate's level of performance?
- Is the precision of proficiency estimates consistent across the range of theta? Does an overall estimate of "reliability" suggest there is sufficient precision in the overall scores to distinguish among candidate performances?

The unidimensional PCM was fit to the 2015 sample of 27,759 candidates. Models were estimated using marginal maximum likelihood as carried out with the "TAM" package in R (Kiefer, Robitzsch, & Wu, 2015), which uses statistical approaches based on those in the software program Conquest (Wu, Adams, Wilson, & Haldane, 2007). As noted above, edTPA scores are derived from the ratings of a single scorer who scores the entire portfolio; rubric scores may reflect some rater effects. Additionally, the results presented below are based upon aggregating data across credential areas. Because edTPA is used based on a single total score calculated equally across fields with 15 rubrics, this analysis provides evidence about how this measure functions overall.

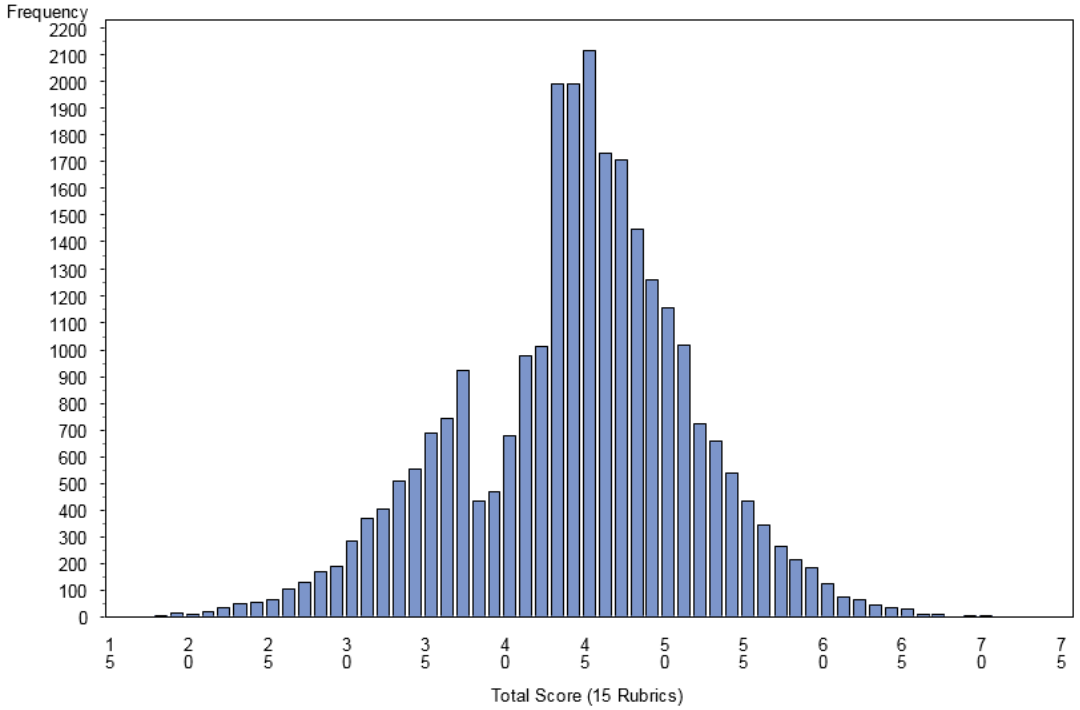
To evaluate fit, INFIT mean square statistics were computed for each rubric and examined to identify rubrics with INFIT values less than 0.75 or greater than 1.33, which would suggest a lack of fit. Plots of expected and observed rubric scores across the theta range were compared across the theta range to identify potential model misfit. A Wright Map depicting the distribution of candidate proficiency estimates alongside rubric threshold parameter estimates was inspected to determine whether: a) 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). 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.** Person separation reliability, similar to Cronbach's alpha, was estimated at 0.910, indicating a high level of consistency.

Candidate Performance

Overall Scores

The following figure presents the score distribution of 27,172 edTPA portfolios, in fields scored based on 15 rubrics and submitted January 1 - December 31, 2015, the second 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 normally distributed across this range. The dip in scores around 37-42 is an artifact of the double scoring process automatically applied to all portfolios that fall within the state or national double-scoring band. 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 are presented in the following table. As a reference, rubrics are listed below by title.[4]

Rubric	Mean	S.D.	Min	Max
Task 1: Planning				
P01	3.1	.7	1	5
P02	3.1	.7	1	5
P03	3.1	.7	1	5
P04	3.0	.7	1	5
P05	3.0	.7	1	5
Task1 Total	**15.3**	**2.8**	**5**	**25**
Task 2: Instruction				
I06	3.1	.5	1	5
I07	3.0	.6	1	5
I08	2.9	.7	1	5
I09	2.9	.8	1	5
I10	2.8	.7	1	5
Task2 Total	14.7	2.5	5	25
Task 3: Assessment				
A11	**3.0**	**.8**	**1**	**5**
A12	**3.1**	**.9**	**1**	**5**
A13	2.6	.8	1	5
A14	2.7	.7	1	5
A15	2.9	.8	1	5
Task3 Total	**14.2**	**3.2**	**5**	**25**
Overall Total	**44.2**	**7.4**	**15**	**75**

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

Task 1: Planning

P01. Planning for Content Understandings
P02. Planning to Support Varied Student 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 Use of Feedback;
A14. Analyzing Students' Language Use and Content Learning
A15. Using Assessment to Inform Instruction

[4] Descriptive statistics for Task 4 rubrics of the Elementary Education Handbook (M19: Analyzing Whole Class Understandings, M20: Analyzing Individual Student Work Samples, M21: Using Evidence to Reflect on Teaching) are presented in Appendix C.

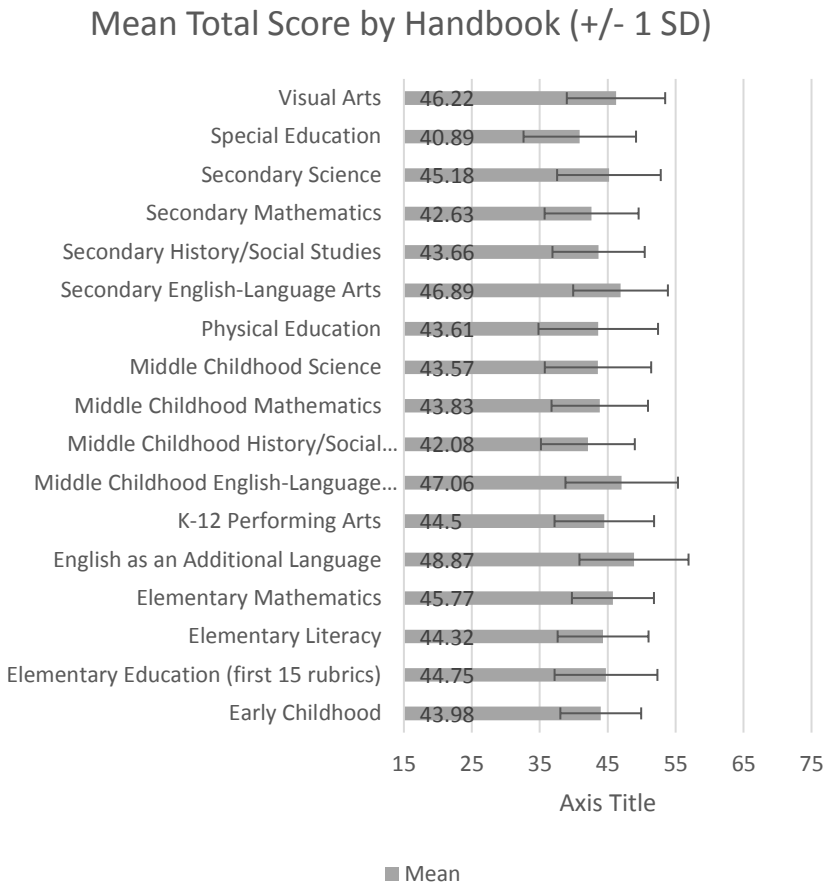
Descriptive Summary by Task and Rubric

The average edTPA score across 27,172 portfolios from fields with 15-rubric handbooks was 44.2, with a standard deviation of 7.4. This average performance shows growth from the 2013 field test data where the average score was 42.8 (SD = 8.17), and a slight decrease in performance but a smaller standard deviation from the 2014 Administration Report where the average score was 44.3 (SD=7.8). Scores ranged across the entire range of possible scores, from 15 to 75. These findings parallel those from the 2013 field test, and the 2014 operational year showing that candidates performed most highly on the planning task, followed by the instruction task, and then the assessment task. This is also consistent with other studies and literature in teacher education that identifies the evaluation and response to students' learning as one of the more challenging elements of teaching (Black & William, 1998; Mertler, 2009). Based on the national recommended cut score of 42, the pass rate for candidates who submitted an edTPA portfolio in 2015 was 71% across all states, and 72% in states using the assessment consequentially. Note that to date, passing scores established by states range from scores of 35 to 41.

Performance by Content Field

The following graph shows total score means by subject area field for edTPA portfolios submitted January 1 - December 31, 2015, in fields scored based on 15 rubrics. Data reflect complete submissions in fields with sample size (N) > 100. For double-scored portfolios, the average score across the two was used. Bars represent scores one standard deviation (SD) below and one SD above the mean. Mean total scores for low incidence fields can be found in Appendix C.

Tables in Appendices D and E provide mean candidate performance, an abbreviated distribution of total scores for national fields, and distributions of rubric-level scores and condition codes reported by field. **Due to differences in sample size, content knowledge demands, and low numbers of submissions in some fields, comparisons across fields should be approached with caution.**



Bars represent scores one standard deviation (SD) below and one SD above the mean.

All edTPA handbooks examine the same underlying constructs and follow the same architecture with 80% overlap in content, with particular subject-specific elements that align to standards and expectations for pedagogy and student learning in that field accounting for the other 20%. 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 with appropriate sample size.
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 with appropriate sample size.
4. Scoring trainers and supervisors are consulted to confirm scoring and backreading processes and flag any recurring questions from scorers.
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 email to SCALE or AACTE staff are reviewed.
7. Review of and clarification to handbooks, scorer training, and support materials is conducted annually based on all quantitative and qualitative data.

Special Education Performance Examined

Based on requests from the field, a deep investigation into the score performance in the field of Special Education was conducted in 2015. Data on performance across different subject fields indicates that the scores of candidates taking edTPA in Special Education are somewhat lower than those in other high incidence fields (mean = 40.9). To examine this outcome we explored many factors that might help interpret the candidate performance including preexisting differences in the candidates going into the field, in requirements and standards of the field, in handbooks, in scorer consistency, in program curricula and structure, and/or in the demands and challenges inherent in serving this widely diverse student population. Of course these systems and causal mechanisms are likely to be interrelated. SCALE took a multipronged approach to investigate this trend and potential contributing factors:

- **Inter-rater reliability:** Analyses of randomly double-scored Special Education portfolios indicate that agreement rates between independent scorers and Kappa N estimates meet standards of total agreement > 90%, and kappa n > .80. Reliability data for each rubric by field is used to inform scorer accuracy as well as communication with trainers, and supervisors to guide new scorer training revisions.
- **Differential item analyses**: Analyses were run to examine systematic differences in rubric difficulty for candidates with same total scores. These analyses confirmed that scores were systematically lower in Special Education across all rubrics. The rubrics with the largest differences when compared to scores in other fields were rubrics requiring the candidate to attend to two learning targets for Special Education students (rubrics 1, 2, 3, 5, 11, and 15.)
- **Comparisons of performance patterns across all content fields:** Analyses were conducted to test whether rubrics appeared to be differentially harder (or easier) for candidates taking edTPA in Special Education. These analyses indicated that candidates taking edTPA in Special Education did not systematically earn higher or lower scores by rubric, when compared to other candidates with the same total score in other fields.
- **Breakdown of demographic subgroups represented within field:** The breakdown of candidates by gender, ethnicity, teaching context, primary language, and level of education of candidates taking

Special Education edTPA are comparable to that in other fields. In other words, the pattern of lower scores in Special Education cannot be attributed to the under- or over-representation of any particular subgroup within the pool of candidates taking edTPA in this field.

- **Review of differences within field placements across programs and states:** Differences in policy, preparation of candidates, structure of the field, and approach to edTPA implementation all contribute to how candidates score within and across fields. The pattern of performance seen nationally does not represent that of every state or every program; in some programs, scores on the Special Education edTPA are equal to or exceed the mean for all fields.
- **Feedback, review and input from:**
    - State Technical Advisory Committees (NY, OH, CA, and WA)
    - Scorers, trainers, scoring supervisors via survey and the Online Community
    - National User Group/Design Team, key state leads group, state advisory groups and edTPA coordinators and the National Policy Advisory Board
    - Subject-specific design teams; scoring supervisors, trainers, and scorers; a Council for Exceptional Children (CEC) advisory group of special education experts; and comments, questions, and suggestions indicating areas of confusion from faculty and candidates
    - A group of Georgia special educators convened for orientation to the edTPA handbook by The Collaboration for Effective Educator Development, Accountability, and Reform (CEEDAR)
    - A committee of special educators convened by the state of New York

These investigations supported the claim that the edTPA Special Education Handbook assesses constructs relevant to, and aligned with, the standards of the profession, and that it meets the reliability and validity criteria put forward by the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014). Based on these data and sources of feedback there were areas of the Handbook that should be modified to address the comments from the expert review of the handbooks to clarify directions and understandings. The edTPA Special Education design team made the following revisions to the Special Education Handbook for 2015-16 based on the review above:

- Change from two learning targets to one learning goal plus planned support.
- Change from breaking down expressive/receptive communication skill into subskills to a focus on support for focus learner use of the expressive/receptive communication skill to participate in learning tasks and/or to demonstrate learning.
- Work sample chosen to illustrate analysis and feedback in Assessment Task 3, changed from the final assessment to any assessment during the learning segment.
- Some rubrics modified in line with minor generic rubric changes made to all handbooks.

The handbook changes noted above went into effect in Fall 2015 and the dataset presented in this report includes both Spring 2015 (former handbook) and Fall 2015 (new handbook). Our most recent national data for the 2015-2016 academy year and the first full year of candidates using the new handbook reveals an increase in performance with an overall mean of 42.8.

Performance by Consequential Use

edTPA portfolios officially scored in the 2015 operational year represent submissions from candidates in 27 states. Of the 27,759 portfolios scored, 5,720 were submitted by candidates in states that do not currently have policy for edTPA use, and 21,452 were submitted in states with consequential policy. States without policy with submissions in 2015 are OH, NC, CO, WV,

WY, NJ, PA, AL, MD, OK, DE, UT, AR, IN, OR, AZ, HI, and VA. States with consequential policy and submissions in 2015 are NY, GA, IL, MN, WA, WI, TN, CA, and IA. Appendix E presents the approximate percentage of portfolios coming from each state. 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. The table below 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.

As predicted, edTPA scores were higher in states with policy requiring edTPA, with non-policy states having a mean of 43.15 and policy states having a mean of 44.53. 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. This pattern is present across most content fields (see Appendix F), although low sample sizes in some fields mean that any interpretations or comparisons should be approached with caution at this time. Typically, faculty preparing candidates in states with consequential policy have had more time to become familiar with and utilize edTPA as an assessment and educative tool, as well as to draw upon edTPA resources and build supports for their candidates. It is also an artifact of high-stakes assessment that higher stakes influence higher levels of effort, motivation, and consistency for all stakeholders. Ongoing integration of edTPA by states and EPPs will inform research into the approaches and practices that best facilitate, support, and assess teaching effectiveness of pre-service teaching candidates.

	States without Policy			States with Policy		
	N	Mean	Std. Deviation	N	Mean	Std. Deviation
Task 1: Planning						
R01	5720	3.105	.7039	21452	3.147	.6923
R02	5720	3.013	.7664	21452	3.077	.7445
R03	5720	2.998	.7038	21452	3.096	.7006
R04	5720	2.95	.662	21452	3.04	.679
R05	5720	2.954	.7549	21452	3.013	.7346
Task 1 Total	5720	15.0242	2.85065	21452	15.3769	2.78426
Task 2: Instruction						
R06	5720	3.105	.4896	21452	3.141	.4962
R07	5720	2.948	.6370	21452	2.993	.6254
R08	5720	2.884	.6877	21452	2.958	.6834
R09	5720	2.755	.7895	21452	2.915	.7397
R10	5720	2.685	.6834	21452	2.799	.6942
Task 2 Total	5720	14.3770	2.47015	21452	14.8054	2.44296
Task 3: Assessment						
R11	5720	2.853	.8504	21452	2.984	.8384
R12	5720	2.964	.8826	21452	3.085	.8487
R13	5720	2.448	.7867	21452	2.590	.7830
R14	5720	2.66	.726	21452	2.74	.745
R15	5720	2.769	.8262	21452	2.903	.8074
Task 3 Total	5720	13.6942	3.24139	21452	14.2995	3.17362
Total Score National	**5720**	**43.15**	**7.532**	**21452**	**44.53**	**7.365**

Performance by Demographic Subgroups

When submitting an edTPA portfolio for official scoring, the candidate is asked to provide demographic information in several categories: gender, ethnicity, teaching placement context, education level, and primary language. Analyses of performance by subgroup within these categories included only portfolios submitted in states that have policy for consequential use of edTPA. In states without such policy, many factors may affect candidate performance into the assessment of teaching competence such as variability in the level of implementation, support structures, level of effort, and candidate motivation and preparation. The portfolios represented here were submitted in CA, GA, IA IL, MN, NY, TN, WA, and WI.

The analyses revealed small differences in performance across some of the subgroups. **It is important to note the difference in sample sizes of some of the subgroups within each demographic category may affect the ability to generalize these results to the national pre-service teaching 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, and other factors that may contribute to a candidate's score and cause differences in performance. What follows is a description of subgroup performance in the following categories: Teaching Context, 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

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 settings had the highest average scores, while candidates teaching in rural settings had the lowest average scores. For the ANOVA and Games-Howell post hoc analyses, see Appendix G.

Teaching Context	N	Mean	Std. Deviation
Rural	3769	43.29	7.39
Rural/Suburban	1843	43.35	7.52
Suburban	7130	45.26	7.11
Suburban/Urban	2168	44.18	7.35
Urban	6542	44.90	7.45

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 rural settings have significantly lower average scores.

Different teaching contexts present different sets of experiences and opportunities for a pre-service teacher candidate. 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 reflect on how they serve and prepare their candidates and to scaffold conversation about teaching strategies that best support learners across various teaching contexts.

Ethnicity

Data from the 2015 operational year indicated that the large majority of candidates submitting edTPA portfolios were White (76.79%), followed by

Hispanic (6.25%), African American (5.6%), Asian (4.12%), and American Indian or Alaskan (.29%), with 2.61% identifying as Multiracial, 1.29% Other, and 3.04% not identifying ethnicity. **The disproportionate representation of White candidates and the relative small sample sizes of other groups must be considered when making comparisons or generalizations to other samples or to the general population of teacher candidates.**

The table below shows the sample size, average scores, and standard deviations of each subgroup. For the ANOVA and Games-Howell post hoc analyses of these results, see Appendix G.

ETHNICITY	Mean	N	Std. Deviation
African American/Black	41.07	1202	7.730
American Indian or Alaskan Native	43.65	62	5.959
Asian or Pacific Islander	45.46	884	6.874
Hispanic	44.69	1341	7.363
White	44.69	16473	7.264
Multi	45.02	560	7.41
Other	44.57	277	7.66
Undec	44.98	653	8.14

Analyses revealed that there was no significant difference between the average scores of White candidates and Hispanic candidates. While the average score of African American candidates was lower than those of other subgroups ($p < .01$), the fact that African American candidates made up a very small portion of the candidate pool (5.6%) should be noted.

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 subgroup than between the subgroups. The difference in means between the White and African American subgroups is 3.62 points, and the pooled standard deviation (7.26+7.73)/2= 7.57. The difference in the mean performance of African American and White candidates in this sample, then, is about one half (.48) of a standard deviation (3.62/7.57). These findings contextualize the magnitude of the difference and demonstrate that the scores of candidates in these two subgroups overlap substantially. Placing this finding in the context of assessment of teacher education, the gap between average scores of White candidates and other subgroups is smaller than that seen with other more traditional standardized assessments of initial teaching (e.g., Goldhaber and Hansen, 2010).

The performance of candidates was also examined by subgroup within each teaching context to see whether the pattern of the overall sample was consistently across the different placements (contexts). This examination revealed that the difference in means of the White and Hispanic candidates were consistently less than 1 point within all teaching contexts. There was greater variation in mean differences between African American and White candidates, with differences being greater in suburban/urban settings than in rural settings or urban settings. As noted in the 'Teaching Context' section above, the performance of candidates in rural settings is systematically lower than that in all other contexts, suggesting that further research in this area is needed.

Placing this finding in the context of assessment of teacher education, the gap between average scores of White candidates and other subgroups is smaller than that seen with other more traditional standardized assessments of initial teaching.

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. While caution must be taken in making generalizations based on such small sample sizes, these findings are consistent with other reported portfolio-based performance assessment data (e.g., NBPTS, PACT, ProTeach). As more data become available, additional research is planned at the state and national levels – we are committed to supporting research to better understand these differences in performance.

Primary Language

Candidates were asked to identify whether English is their primary language. Primary Language English candidates scored higher than those candidates who indicated their Primary Language was not English; this difference, while small (.88) was statistically significant ($p < .01$).

PRIMARY LANGUAGE	N	Mean	Std. Deviation
English	20647	44.55	7.37
Other	561	43.67	7.17

While the disproportionate N size for these two populations would warrant caution in overgeneralizing (candidates who indicated their Primary Language was not English represent 3.6% of the population), however the small differences in mean performance is an encouraging finding.

Gender

In this sample, 77.6% of submissions came from female candidates, and 22.3% from male candidates. Female candidates scored higher than their male counterparts; this difference (1.47), was statistically significant ($p < .01$).

GENDER	N	Mean	Std. Deviation
Male	4743	43.39	7.75
Female	16452	44.86	7.21

Follow up analyses reveal that the difference was greatest in suburban/urban teaching contexts (2.02 points), and smallest within urban contexts (1.24 points); see Appendix H. These findings suggest that the difference in performance by gender may vary based on other variables such as educational background, or preparation program.

Education Level

The achieved level of education prior to taking edTPA was reported by candidates. Candidates holding a doctorate degree had the highest average scores; due to low sample size this subgroup was not included in statistical comparisons of mean difference. For the ANOVA and Games-Howell post hoc analyses, see Appendix G.

EDUCATION LEVEL	N	Mean	Std. Deviation
High school/Some college	10825	44.07	7.12
Bachelor's/Bachelor's plus credits	8777	45.18	7.48
Master's/Master's plus credits	1782	44.09	7.99
Doctorate	68	46.29	8.03

Due to the significant disparities in the size between the Masters/Master's plus credits sample and that of the HS/Some college and the Bachelor's/Bachelor's plus credits samples, results should be interpreted with caution. One hypothesis is that candidates who take edTPA after earning a Master's degree may have a background in a different field or have had less coursework and/or student teaching experience prior to taking edTPA. 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.

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, 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 a finding that a factor is a "statistically significant" predictor does not necessarily indicate that this factor makes a substantial or meaningful difference in outcomes. The percent of variance explained (Delta 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, $F(21, 21430) = 42.77$ ($p<0.01$), 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 (%)
School Context	1.15
Ethnicity	1.48
Gender	0.65
Education Level	0.68
Primary Language	0.09
Overall Model	4.02

Overall, this model accounts for only 4.02% of the variance in total scores (R2 = .04); 95.98% of the variability in scores is explained by other factors not accounted for by the variables included in this model. This result highlights that demographic factors account for a very small portion of the variables that contribute to how a candidate scores on their edTPA. In other words, a candidate's demographic characteristics alone are a poor predictor of a candidate's edTPA performance or readiness to teach. This finding further supports the conclusion that while some statistically significant differences do exist between subgroups, approximately 96% 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 the 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.

...a candidate's demographic characteristics alone are a poor predictor of a candidate's edTPA performance or readiness to teach.

Reliability Evidence

Inter-rater agreement

The table below shows inter-rater agreement for the 2015 edTPA administration cycle (January 1, 2015 - December 31, 2015). 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
- Total agreement – proportion of cases in which the pairs of scores are ± 1 score point apart from each other.

The data set included, 2228 complete submissions (approximately 10% of the total number of examinees) that were scored independently by two scorers as part of the random sample of double-scored portfolios for the 2015 administration cycle. Across all 15 rubrics and 2,228 candidates, independent scorers assigned the same score (exact agreement) in approximately 53.2% of all cases. Also, scorers assigned scores that were one point apart (adjacent agreement) in approximately 41.5% of all instances. When combining exact and adjacent agreement into a total agreement, scorers assigned scores that were the same or +/- 1 point apart in approximately 94.7% of all cases. These exact and adjacent agreement rates are consistent with that of other performance assessments, such as the NBPTS.

The kappa n provides chance-corrected total agreement, or inter-rater agreement measures that result from removing total agreement that may have occurred randomly (Brennan & Prediger, 1981). Chance-corrected agreement ranges from 0 to 1. There are no widely accepted guidelines for what constitutes an adequate value of the coefficients, although higher values represent greater levels of agreement. Table 2 shows kappa-n ranged from 0.839 (rubric 2 and rubric 12) to 0.95 (rubric 6), with an average value of 0.89. This outcome corroborates that scorers tend to assign scores within +/- 1 and rarely assign scores that differ by more than 1 point. The overall chance-corrected total agreement rate (0.89) is consistent in magnitude with the kappa n rate found in the 2014 Operational Year (0.86)

Task	Rubric	Inter-Rater Agreement			
		Exact	Adjacent	Total	Kappa N
Task 1: Planning	Rubric 01	0.539	0.412	0.951	0.897
	Rubric 02	0.488	0.434	0.923	0.839
	Rubric 03	0.545	0.414	0.959	0.915
	Rubric 04	0.542	0.412	0.954	0.905
	Rubric 05	0.517	0.428	0.945	0.885
Task 2: Instruction	Rubric 06	0.676	0.300	0.976	0.950
	Rubric 07	0.563	0.404	0.967	0.932
	Rubric 08	0.522	0.425	0.948	0.892
	Rubric 09	0.531	0.403	0.934	0.862
	Rubric 10	0.526	0.428	0.954	0.904
Task 3: Assessment	Rubric 11	0.513	0.422	0.935	0.864
	Rubric 12	0.485	0.438	0.923	0.839
	Rubric 13	0.502	0.430	0.933	0.860
	Rubric 14	0.538	0.423	0.961	0.919
	Rubric 15	0.491	0.452	0.943	0.882
Overall	**Average**	0.532	0.415	0.947	0.890

Internal Consistency

Cronbach's alpha is a measure of internal consistency of raw test scores, an important characteristic of test scores that indicates the extent to which the items of the assessment measure the intended common construct (Cronbach, 1951). Cronbach's alpha estimates range from zero to one, and higher values reflect higher levels of consistency of a person's scores across the items (rubrics).

The table below shows edTPA estimates of Cronbach's alpha coefficient for the 2015 administration cycle. The table shows descriptive statistics for total scores and reliability estimates for individual fields and the overall group. The data set included 27165 complete submissions (excluding portfolios with condition codes). The estimation of reliability took place with the data from the first rater. Reliability coefficients ranged from 0.852 (Family and Consumer Sciences) to 0.934 (Library Specialist), with an overall alpha of 0.907, indicating a high level of consistency across the rubrics, meaning that the rubrics as a group are measuring a common construct of teacher readiness.

The person separation reliability calculated as part of the IRT internal structure analyses presented in the *'Validity'* section of this report was estimated as 0.910, indicating a high level of reliability for distinguishing among candidates' levels of performance. This index is similar to Cronbach's alpha. Generally, values of 0.90 or greater are expected for such reliability indices.

Field Name	N	Mean	Variance	Cronbach's alpha
Agricultural Education	106	48.057	47.063	0.918
Business Education	109	46.550	48.676	0.914
Early Childhood	2631	43.979	35.450	0.875
Elementary Education	3958	44.752	57.359	0.914
Elementary Literacy	3161	44.320	44.742	0.919
Elementary Mathematics	2530	45.771	36.541	0.889
English as an Additional Language	417	48.873	64.414	0.924
Family and Consumer Sciences	64	41.125	33.00	0.852
Health Education	126	34.079	55.242	0.930
K-12 Performing Arts	1325	44.500	53.690	0.914
Library Specialist	49	46.347	125.315	0.934
Middle Childhood English-Language Arts	426	47.056	68.482	0.919
Middle Childhood History/Social Studies	344	42.084	47.803	0.897
Middle Childhood Mathematics	514	43.829	50.388	0.893
Middle Childhood Science	382	43.568	61.306	0.905
Physical Education	866	43.614	77.490	0.919
Secondary English- Language Arts	1879	46.886	48.586	0.915
Secondary History/Social Studies	1786	43.657	46.105	0.910
Secondary Mathematics	1547	42.628	47.944	0.893
Secondary Science	1248	45.179	58.218	0.912
Special Education	3043	40.889	68.536	0.923
Technology and Engineering Education	51	42.196	75.761	0.881
Visual Arts	603	46.217	52.449	0.895
Overall	**27165**	**44.181**	**56.720**	**0.907**

Setting Cut Scores Using Standard Error of Measurement

In assessment, each time an examinee takes a test there is a random chance that the score will be slightly different, and applying the standard error of measurement (SEM) is one way to take this into account. The SEM allows educational analysts to determine the range of scores an examinee would receive if tested repeatedly without studying or contemplating the answers between tests. By applying this technical adjustment, a given examinee's score may be more representative of "true" knowledge because the variation in scores is taken into account, and it provides a safeguard against placing undue emphasis on a single test score.

There are different ways to estimate the standard error of measurement. For edTPA we used a method based on the total number of score points available (75) and the recommended passing standard (Lord, 1959; Gardner, 1970). In determining state-specific cutscores for edTPA, state agencies are provided with the panel-recommended passing standards along with SEM adjustments so that they may consider the impact on pass rates overall or by subgroup for scores at a given SEM adjustment. Providing these SEM considerations gives states context for a number of policy considerations involved in determining a passing standard for a consequential assessment in a state. See the 2014 Administration Report for a full description of the SEM process.

Candidate Passing Rates

The following table reports the percent of candidates (out of 27,172) who would have "passed" edTPA (based on the edTPA 2015 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 (within one and a half standard error of measurement of the Recommended Professional Performance Standard). 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.

	Candidate Passing Rates
	Overall Passing Rate
Cut Score	
35	88.9%
36	86.4%
37	83.6%
38	80.2%
39	78.6%
40	76.9%
41	74.4%
42	70.8%

State Standard Setting

edTPA Standard Setting Event Overview

edTPA state standard setting conferences occur over one or multiple days. The method used to conduct the standard setting is the Briefing Book Method (Haertel, Beimers, & Miles, 2012). The Briefing Book Method (BBM) is an evidence-based standard setting method intended to develop an appropriate and defensible cut score that can be supported with a validity argument. The BBM provides a framework and approach to standard setting rather than a specific set of steps or procedures that must be followed exactly. The primary aim is to follow a process that allows a body with the appropriate authority and knowledge to reach a defensible and appropriate judgment of a passing cut score.

Participants in the conference include groups of subject area experts, educators, and policymakers who are convened into a panel for the standard setting session. For each participant group, the conference organizers strive to have an equal mix of higher education faculty, non-traditional educational preparation program providers (e.g., area education service organizations), and P-12 educators. Panelists are informed of the purpose of the assessment and are provided with the "briefing book" to guide their activity. Prior to the meeting, each invited panelist receives edTPA handbooks, rubrics, scoring materials, and three previously scored sample portfolio submissions representing different performance levels across various content areas. Panelists are asked to review materials submitted by candidates and the scoring evidence identified by trained benchmarkers for the submissions assigned to them. During the facilitated session, panelists familiarize themselves with the assessment and with the information contained in the briefing book. After a series of "Policy Capture Activities" examining whole portfolios and score profiles representing a range of candidate performances, panelists recommend an initial cut score (which may also be referred to as a "passing standard") for each task, which is then discussed and evaluated based on impact data. Following that, panelists recommend a final cut score.

edTPA Guiding Question

Throughout the standard setting event and examination of sample edTPA score profiles, a prompt and a guiding question are used and revisited to frame all discussions. This contextual prompt and guiding question provide a common framework in which all participants anchor their decisions.

- Think about a teacher candidate who is just at the level of knowledge and skills required to perform effectively the job of a new teacher in (Insert State Name) public schools.
- Guiding question: What score (the sum of all of the rubric scores of edTPA) represents the level of performance that would be achieved by this individual?

The purpose of the edTPA standard setting guiding question and contextual prompt is to identify the performance expectation of an initially licensed, classroom-ready teacher. The step-by-step standard setting process of examining actual candidate submissions, candidate score profiles, and impact data guides participants to determine the candidate performance on edTPA that, as stated in the Briefing Book Method, "just meets the definition of performing effectively the job of a new teacher."

Refer to the 2014 edTPA Administration Report for in-depth explanation of the steps included in the standard setting process.

Outcomes

Typically, in setting a cut score for a pass-fail decision, a standard error of measurement is applied to the recommended score so as to reduce decisions influenced by measurement error (e.g., false negatives). The full standard error of measurement puts a lower bound on the recommended score of about five points.

States may set their own passing scores based on state standard setting panels that take into account state-specific data, measurement data, and the state's policy considerations. As discussed by the national standard setting panel members, as well as the state panelists, states may consider setting

their initial cut score lower than the panel-recommended score to give programs time to learn to deliver and support edTPA activities and to support candidates' preparation of their submissions. This “phase-in” strategy allows for a ramping up of the state based standard over time, eventually reaching the panel-recommended score, or other cut score, after a defined period of time. An example of a phase-in strategy would be to establish a passing score at -1 SEM from the panel-recommended score, raise the passing score to -1/2 SEM after the first year of operational use, and finally to raise the passing score to the panel-recommended score after the second year of operational use. This allows states time to examine operational data during the defined timeframes, and review pass rates over time. As warranted, the state performance standard can be reviewed and adjusted as appropriate over time.

State Based Passing Standards

Between fall 2013 and the end of 2015, the following states established state-based passing standards as follows (for 15-rubric fields):

- Alabama (37)*
- Arkansas (37)*
- California (41)
- Delaware (38)*
- Georgia (35)
- Illinois (35)*
- Iowa (41)
- Minnesota (Task 1: 13, Task 2: 13, Task 3: 12)
- New York (41)
- Washington (35, excludes Student Voice)

  *indicates states utilizing a methodology outside of the Briefing Book.

Note that state-based passing standards may be reevaluated and adjusted, as driven by state reviews. The passing standards cited above were in use during the 2015 calendar year, the date range which this report covers. See the edTPA State Policies Overview for the most up to date information on state policies and any established state consequential score information, including planned adjustments over time.

TAC recommendations for future directions

The edTPA National Technical Advisory Committee (TAC; for members, see list in Appendix I) has reviewed the evidence presented in this report; their input guided the analyses and interpretations presented. The discussion included planned and recommended future directions that will add to the validity evidence outlined here and inform state and program policy about the role of edTPA in the education of their teacher candidates. The diversity of expertise and perspectives represented by the TAC provided for rich discussion and suggestions for additional analyses and research questions, which are represented throughout the report. The TAC will continue to provide advice and counsel to the nationally convened Research Group as they review the research literature on edTPA and conduct new studies of consequential impact and predictive validity.

Conclusion

edTPA was developed for the profession by the profession to be a reflection of the broad skills and competencies necessary to be a successful teacher. Founded on the subject-specific architecture of the National Board for Professional Teaching Standards’ assessments and the work in California on the Performance Assessment of California Teachers (PACT), edTPA is aligned with the Interstate Teacher Assessment and Support Consortium (InTASC) standards for beginning teacher licensing (2013). The development of edTPA, content validity studies and subsequent revisions, and job analyses add to research-based evidence of effective teacher performance and capture the skills, knowledge, and abilities of a novice teacher. The Review of Research on Teacher Education presents the research foundation of edTPA as an assessment of teacher readiness as defined by the leading experts and

existing literature on teacher preparation. As with the field test data, data from the first two years of operational use are consistent with *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), and affirm the reliability and validity evidence necessary for edTPA to be used for the evaluation of teacher candidates.

The use of edTPA scores by EPPs and state agencies as a reflection of a candidate's readiness to be an effective and proficient educator is predicated on observed scores being accurate, unbiased, reliable, and consistent across relevant conditions of measurement. The scoring model, training design, double scoring and adjudication processes, and quality management of scorers describe the rigorous scoring model applied to the reporting of edTPA final scores, and analyses of interrater reliability quantify the precision and reliability of these scores. The confirmatory factor and partial credit model analyses of internal structure support the construction 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' performance. Data on candidates' performance by content field and demographic categories presented in the report suggest that these factors explain a very small portion of variance in total scores, and do not suggest systematic bias against any group or field. As more data become available, the interactions among variables that contribute to candidates' performance on edTPA will inform the use and interpretation of rubric, task, and total scores.

edTPA was designed as a support and assessment system for teachers entering the profession. The use of edTPA to inform decisions about a candidate's readiness to successfully begin his or her career as a teacher is supported by studies that have explored relationships between PACT or edTPA scores with other performance measures of teacher candidates. Summarized in the "Validity" section of this report, emerging studies indicate that performance on these teacher performance assessments is related to candidate performance or readiness to teach: candidates' GPAs, scores on assessments of pedagogy, supervisors' predictions of success, and evidence of student learning. Most importantly, edTPA is an educative assessment that supports candidate learning and preparation program improvement. This report synthesizes the systems of support and resources available to candidates, faculty, and programs; the process of taking the assessment, using it to reflect on individual and program practices, and to use data in systematic and reflective ways. Qualitative and quantitative analyses presented in this report describe the impact of edTPA on programs, faculty, and teacher candidates' educative experience.

More evidence of the concurrent, predictive, and consequential validity of edTPA is eagerly anticipated as data become available; existing research provides strong support that completing edTPA is an educative experience that further improves readiness to teach, while passing edTPA is a signal of readiness that is linked to becoming a more proficient teacher. As more states and educator preparation programs move toward integrated and consistent methods of assessing teacher candidates, it is crucial to continue the examination of reliability and validity arguments of assessments used for licensure/certification, program improvement, and/or program completion. Access to data on candidate performance allows for examination of the preparedness of teachers entering the profession across various skills and constructs. As a subject-specific assessment, edTPA data allows us to consider candidates' readiness to teach for each content field, as well as to present programs with national data trends that in turn inform program preparation and reflection. In collaboration with the edTPA Technical Advisory Committee and the edTPA Research Group, SCALE is committed to continuing research that informs and advances the field of teacher preparation. The findings presented in this report can guide and support educator preparation programs, states, and P-12 partners to inform and reform teaching and learning. It also serves as a call for further research and lays the foundation for research questions that will continue to improve assessment and preparation of readiness to teach P-12 students in every classroom, every school, and every field.

Lastly, as with the case of the National Board for Professional Teaching Standards (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 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 is instrumental to its power as an educative tool. The extensive 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 formative 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 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

Table 1: Confirmatory Factor Analyses: Standardized Factor Loading Estimates

The table below presents the estimated standardized factor loadings for the 1 and 3-factor models in the full sample of portfolios (N=27,759).

1-Factor Model		3-Factor (Task) Model		
Rubric	**F1**	**Planning**	**Instruction**	**Assessment**
1	0.650	0.715	--	--
2	0.639	0.703	--	--
3	0.664	0.706	--	--
4	0.645	0.686	--	--
5	0.680	0.746	--	--
6	0.496	--	0.600	--
7	0.634	--	0.756	--
8	0.602	--	0.725	--
9	0.559	--	0.658	--
10	0.642	--	0.613	--
11	0.718	--	--	0.761
12	0.621	--	--	0.678
13	0.639	--	--	0.702
14	0.667	--	--	0.697
15	0.702	--	--	0.744

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

All factor loadings for the three-factor solution were positive and statistically significant as anticipated (all standardized loadings were greater than .6 in the 3-factor model).

All but one of the factor loadings for the one-factor solution attained magnitudes of at least 0.50, with just a single rubric (Rubric 6) with a factor loading slightly below that target (0.496).

Table 2: Confirmatory Factor Analyses: Task Factor Correlation Matrix

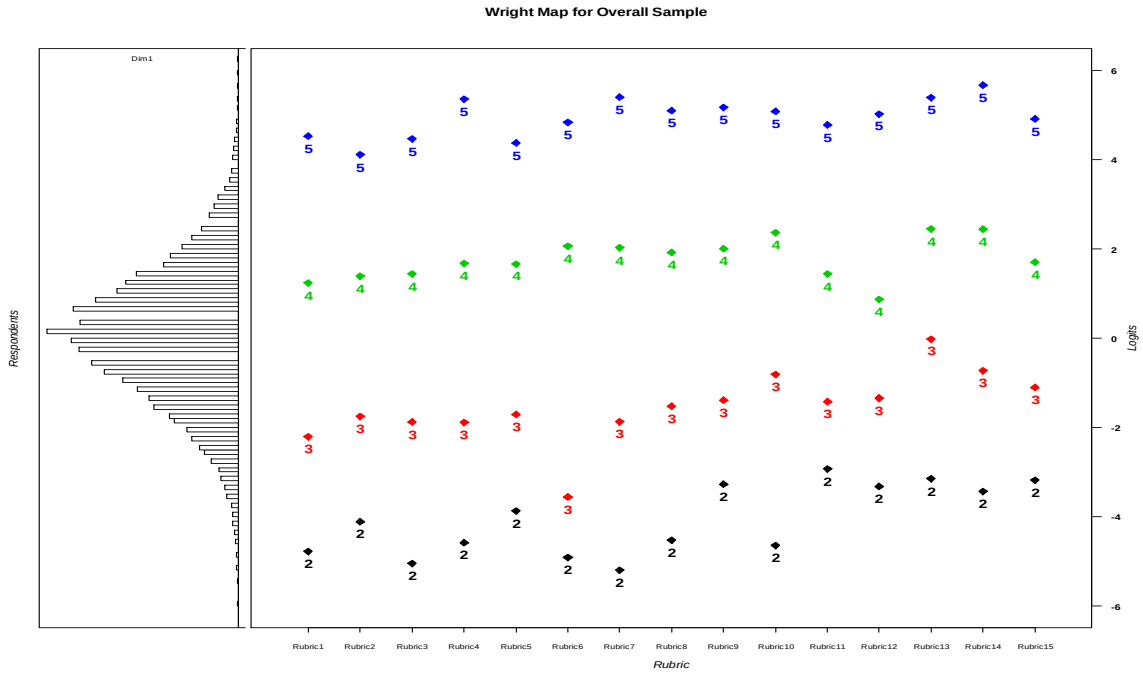
The table below presents the estimated correlations among the task factors in the 3-factor model.

	Planning	Instruction	Assessment
Planning	1.00		
Instruction	0.73	1.00	
Assessment	0.80	0.74	1.00

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.

Table 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. These are a direct function of total scores, which represent estimates of teacher effectiveness. The points on the graph (Thurstonian thresholds) represent the point on the underlying theta 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 theta (logit) 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 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.

Appendix B: Double Scoring Band – Distribution of Scores

Figure 1: Distribution of the first scores

The following figure shows the distribution of the first score on portfolios that are on and around the national cut score. These portfolios were then double scored since they fall within this double scoring band.

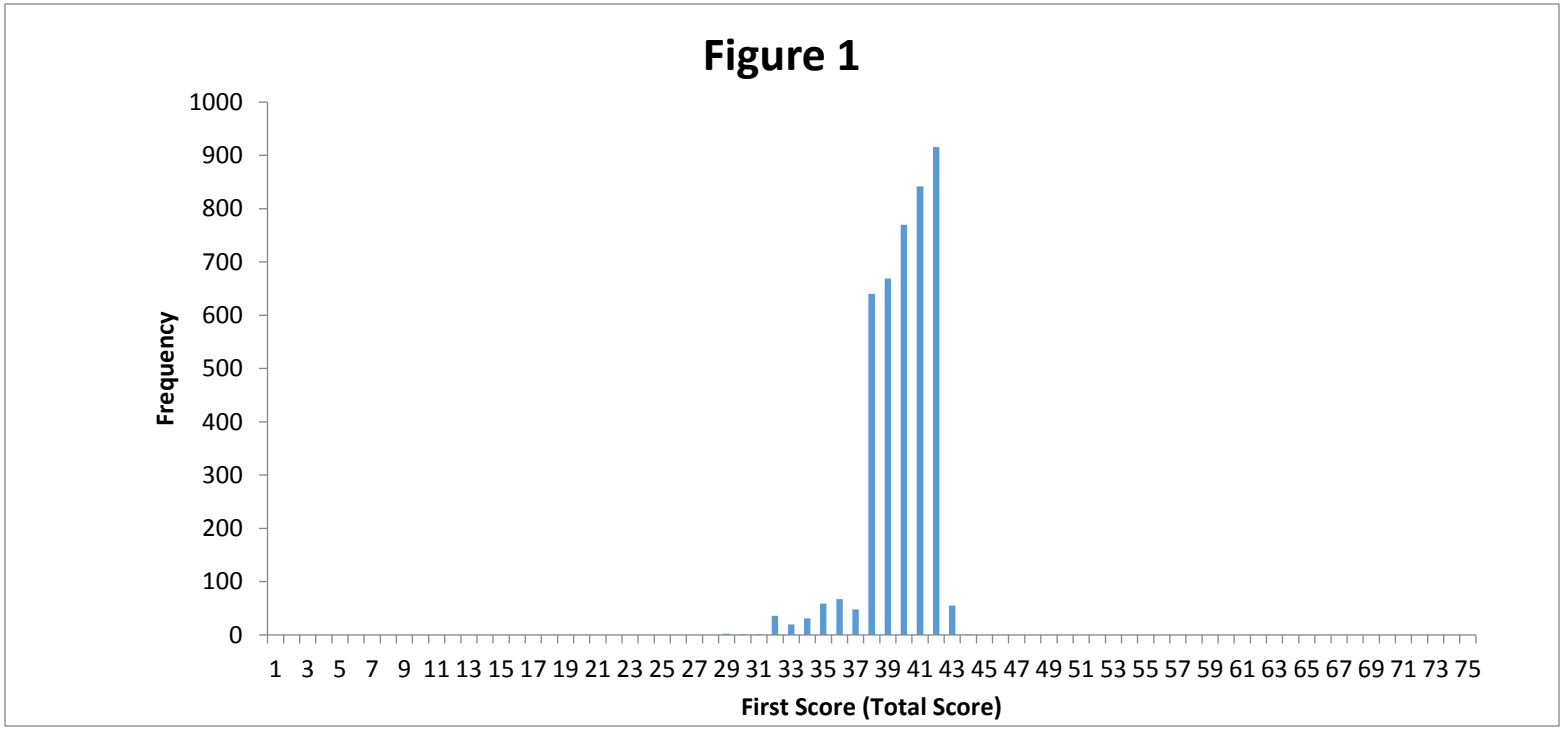
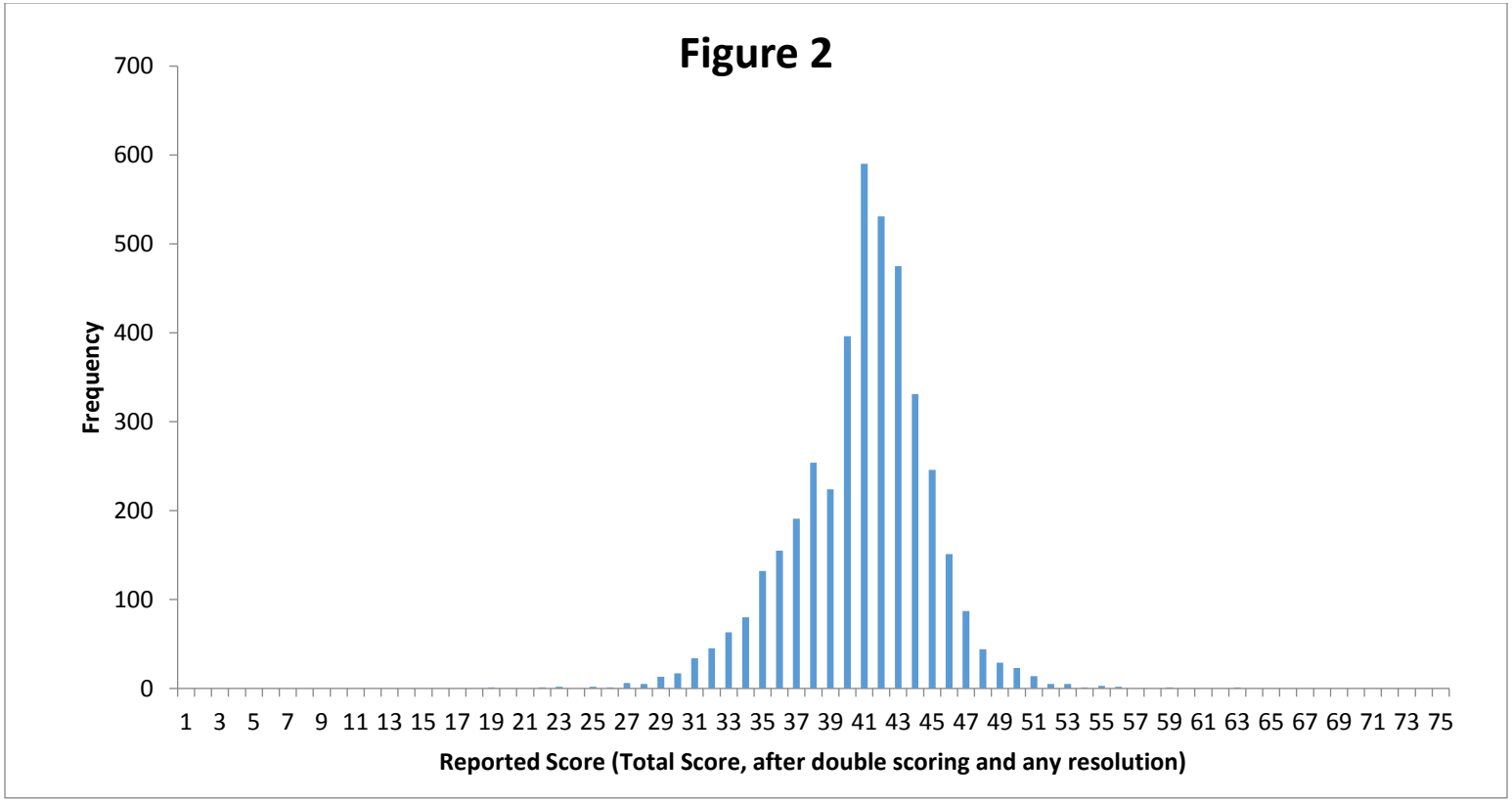


Figure 2: Distribution of the final scores

The following figure shows the final disposition (after double scoring and any resolution) of those portfolios that were within the double scoring band and illustrates the distribution of final scores that were originally around the national cut score.



Appendix C: 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.
- 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.

Means	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	27,172	44.2	3.1	3.1	3.1	3.0	3.0	3.1	3.0	2.9	2.9	2.8	3.0	3.1	2.6	2.7	2.9	15.3	14.7	14.2
Agricultural Education	106	48.1	3.5	3.3	3.2	3.3	3.4	3.4	3.4	3.1	3.3	3.0	3.2	3.2	2.8	2.8	3.0	16.8	16.2	15.0
Business Education	109	46.6	3.3	3.3	3.1	3.1	3.2	3.2	3.2	3.1	3.0	3.1	3.1	3.2	2.8	2.7	3.0	16.2	15.6	14.8
Early Childhood	2,631	44.0	3.2	3.2	3.1	3.1	3.0	3.1	3.0	3.0	2.4	2.8	2.9	2.9	2.4	2.9	2.9	15.6	14.3	14.1
Educational Technology Specialist	1																			
Elementary Education (first 15 rubrics)	3,958	44.8	3.0	3.1	3.1	3.1	3.0	3.1	3.0	3.0	3.0	2.8	3.0	3.1	2.6	2.7	3.0	15.3	14.9	14.5
Elementary Literacy	3,161	44.3	3.0	3.0	3.1	3.0	3.0	3.1	3.0	2.9	3.0	2.8	3.0	3.0	2.7	2.7	3.0	15.1	14.7	14.4

Means	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 Mathematics	2,530	45.8	3.3	3.2	3.2	3.1	3.0	3.1	3.0	3.0	3.0	2.9	3.2	3.2	2.7	2.8	3.0	15.8	15.1	14.9
English as an Additional Language	417	48.9	3.7	3.4	3.4	3.5	3.6	3.4	3.2	3.1	2.8	3.1	3.3	3.3	2.7	3.0	3.3	17.6	15.6	15.6
Family and Consumer Sciences	64	41.1	3.1	3.0	2.9	2.8	2.7	3.0	2.8	2.7	2.8	2.6	2.7	2.6	2.4	2.4	2.5	14.4	14.0	12.7
Health Education	126	34.1	2.2	2.3	2.4	2.4	2.2	2.9	2.2	2.2	2.2	2.3	2.2	2.3	2.0	2.1	2.2	11.5	11.8	10.8
K-12 Performing Arts	1,325	44.5	3.2	3.1	3.1	3.0	3.1	3.0	2.9	2.8	3.0	2.8	3.0	3.1	2.6	2.7	3.0	15.6	14.5	14.3
Library Specialist	49	46.3	3.6	3.6	3.3	3.4	3.4	3.4	3.2	2.9	3.2	2.9	3.0	3.0	2.3	2.5	2.6	17.3	15.6	13.4
Literacy Specialist	6																			
Middle Childhood English-Lang. Arts	426	47.1	3.5	3.3	3.4	3.2	3.3	3.3	3.2	3.1	3.0	2.9	3.1	3.3	2.6	2.7	3.1	16.7	15.4	14.9
Middle Childhood History/Social Studies	344	42.1	3.1	3.0	3.0	2.9	2.9	3.1	2.9	2.7	2.6	2.6	2.8	2.9	2.4	2.4	2.6	14.9	13.9	13.1
Middle Childhood Mathematics	514	43.8	3.1	3.0	3.2	2.9	3.0	3.1	2.8	2.8	3.1	2.7	3.0	3.2	2.6	2.6	2.8	15.1	14.5	14.2
Middle Childhood Science	382	43.6	3.1	3.0	3.1	3.0	3.1	3.2	2.9	2.8	2.7	2.7	2.9	3.2	2.5	2.6	2.7	15.4	14.2	13.9
Physical Education	866	43.6	3.2	3.2	2.9	2.9	2.9	3.2	3.0	3.5	3.1	2.7	2.8	2.9	2.3	2.4	2.6	15.1	15.6	12.9
Secondary English-Language Arts	1,879	46.9	3.4	3.2	3.2	3.2	3.2	3.2	3.2	3.1	3.1	3.0	3.2	3.3	2.7	2.9	3.1	16.2	15.5	15.1
Secondary History/Social Studies	1,786	43.7	3.1	3.0	3.0	3.0	3.0	3.1	3.0	2.9	2.8	2.8	2.9	3.0	2.6	2.7	2.8	15.1	14.5	14.0
Secondary Mathematics	1,547	42.6	3.1	2.7	3.1	2.7	2.8	3.1	2.8	2.8	2.9	2.6	3.0	3.2	2.5	2.6	2.7	14.4	14.2	14.0
Secondary Science	1,248	45.2	3.3	3.2	3.2	3.1	3.3	3.2	2.9	2.9	2.6	2.8	3.1	3.3	2.6	2.8	2.9	15.9	14.4	14.8
Special Education	3,043	40.9	2.8	2.8	2.8	2.9	2.6	3.1	2.9	2.8	2.9	2.5	2.4	2.8	2.4	2.6	2.3	13.9	14.3	12.6
Technology and Engineering Education	51	42.2	3.3	3.0	3.0	3.0	2.8	3.1	2.8	2.7	2.9	2.7	2.8	2.7	2.4	2.5	2.6	15.1	14.1	13.0
Visual Arts	603	46.2	3.5	3.2	3.3	3.1	3.3	3.3	3.1	3.0	3.1	2.8	3.1	3.1	2.4	2.8	2.9	16.5	15.4	14.3

Means	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	587	37.1	3.2	3.1	3.2		3.1	3.2	2.7	2.6	2.2	2.7	2.9	3.0	2.5		2.8	12.6	13.3	11.1
Classical Languages	15	31.7	2.8	2.7	3.0		2.4	2.9	2.2	2.1	1.6	2.2	2.5	2.7	2.3		2.5	10.9	10.9	9.9
World Language	572	37.2	3.2	3.1	3.2		3.1	3.2	2.8	2.6	2.2	2.7	2.9	3.0	2.5		2.8	12.7	13.4	11.1

Means	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	3,869	53.7	3.1	3.1	3.1	3.1	3.0	3.1	3.0	3.0	3.0	2.8	3.0	3.1	2.6	2.7	3.1	2.9	3.0	2.8	15.4	14.9	14.6
Elementary Education	3,869	53.7	3.1	3.1	3.1	3.1	3.0	3.1	3.0	3.0	3.0	2.8	3.0	3.1	2.6	2.7	3.1	2.9	3.0	2.8	15.4	14.9	14.6

Appendix D: Score Distributions by Content Field

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

		N	Mean Score	Distribution of Total Score (%)									
				< 35	35	36	37	38	39	40	41	42	> 42
	All 15-Rubric Handbooks	27,172	44.2	11	3	3	3	2	2	2	4	4	67
	Agricultural Education	106	48.1	4	1	2	2		2	1	2	3	84
	Business Education	109	46.6	3	2	1	2			1	4	3	85
	Early Childhood	2,631	44.0	7	2	3	4	2	2	4	5	4	68
	Educational Technology Specialist	1											
	Elementary Education (first 15 rubrics)	3,958	44.8	12	2	2	2	2	1	2	3	4	69
	Elementary Literacy	3,161	44.3	9	1	3	3	1	2	2	4	4	71
	Elementary Mathematics	2,530	45.8	4	2	2	3	1	1	2	3	4	78
	English as an Additional Language	417	48.9	6	2	1	2	1	1	1	1	1	84
	Family and Consumer Sciences	64	41.1	16	6	2	5		5	3	6	8	50
	Health Education	126	34.1	61	5	3	5	2	4	2	6		13

	K-12 Performing Arts	1,325	44.5	10	2	3	4	2	2	2	4	3	68
	Library Specialist	49	46.3	20	8	2					2		67
	Literacy Specialist	6											
	Middle Childhood English-Lang. Arts	426	47.1	8	2	2	2	1	1	1	2	3	78
	Middle Childhood History/Social Studies	344	42.1	15	4	2	6	3	2	3	5	3	56
	Middle Childhood Mathematics	514	43.8	11	3	5	3	2	2	3	4	4	63
	Middle Childhood Science	382	43.6	14	4	4	6	1	1	1	1	4	63
	Physical Education	866	43.6	17	3	2	4	2	3	3	4	3	61
	Secondary English-Language Arts	1,879	46.9	5	2	2	2	1	1	1	2	3	81
	Secondary History/Social Studies	1,786	43.7	10	2	3	4	1	2	4	5	4	64
	Secondary Mathematics	1,547	42.6	14	4	4	5	3	2	3	5	4	56
	Secondary Science	1,248	45.2	10	2	2	3	1	1	2	3	3	72
	Special Education	3,043	40.9	23	5	4	5	2	3	3	4	4	47
	Technology and Engineering Education	51	42.2	22	2	4	4	2		6		4	57
	Visual Arts	603	46.2	6	1	2	3	1	1	2	3	4	76

		N	Mean Score	Distribution of Total Score (%)							
				< 30	30	31	32	33	34	35	> 35
	All 13-Rubric Handbooks	587	37.1	19	4	4	2	2	4	3	62
	Classical Languages	15	31.7	60					7	7	27
	World Language	572	37.2	18	4	4	2	2	4	3	63

		N	Mean Score	Distribution of Total Score (%)											
				< 40	40	41	42	43	44	45	46	47	48	49	> 49
	All 18-Rubric Handbooks	3,869	53.7	9	1	2	2	2	2	1	1	1	2	2	74
	Elementary Education	3,869	53.7	9	1	2	2	2	2	1	1	1	2	2	74

Appendix E: Portfolios Represented by State

The following table shows all states that submitted edTPA portfolios during the 2015 administrative year, and the approximate percentage of the total sample that each state contributed.

State	Approx %
AL	< 1%
AR	< 1%
AZ	< 1%
CA	4%
CO	1%
DE	< 1%
GA	10%
HI	< 1%
IA	2%
IL	10%
IN	< 1%
MD	< 1%
MN	10%
NC	4%

State	Approx %
NJ	< 1%
NY	25%
OH	12%
OK	< 1%
OR	< 1%
PA	< 1%
TN	4%
UT	< 1%
VA	< 1%
WA	8%
WI	5%
WV	1%
WY	1%

Appendix F: Consequential Use by Content Field

The following table presents 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.

		All			Non-Policy			Policy States		
		N	Mean	Std. Deviation	N	Mean	Std. Deviation	N	Mean	Std. Deviation
	Field									
15 Rubrics	**Agricultural Education**	106	48.06	6.86	35	47.03	8.05	71	48.56	6.19
	Business Education	109	46.55	6.98	18	46.00	4.98	91	46.66	7.32
	Early Childhood	2631	43.98	5.95	1226	43.85	5.98	1405	44.09	5.93
	Educational Technology Specialist	1	39.00	.	.	.	.	1	39.00	.
	Elementary Education (first 15 rubrics)	3958	44.75	7.57	474	43.45	7.83	3484	44.93	7.52
	Elementary Literacy	3161	44.32	6.69	447	43.20	7.45	2714	44.50	6.54
	Elementary Mathematics	2530	45.77	6.04	108	43.55	6.96	2422	45.87	5.98
	English as an Additional Language	417	48.87	8.03	43	50.14	7.26	374	48.73	8.11
	Family and Consumer Sciences	64	41.13	5.74	11	39.82	5.25	53	41.40	5.85
	Health Education	126	34.08	7.43	29	31.83	9.36	97	34.75	6.66
	K-12 Performing Arts	1325	44.50	7.33	299	43.13	7.06	1026	44.90	7.36
	Library Specialist	49	46.35	11.19	5	43.20	17.46	44	46.70	10.50
	Literacy Specialist	6	45.17	6.97	4	47.25	4.57	2	41.00	11.31
	Middle Childhood English-Language Arts	426	47.06	8.28	208	48.43	8.22	218	45.74	8.13
	Middle Childhood History/Social Studies	344	42.08	6.91	158	42.29	6.82	186	41.91	7.01

		All			Non-Policy			Policy States		
		N	Mean	Std. Deviation	N	Mean	Std. Deviation	N	Mean	Std. Deviation
	Middle Childhood Mathematics	514	43.83	7.10	245	44.35	6.91	269	43.35	7.25
	Middle Childhood Science	382	43.57	7.83	172	44.32	7.54	210	42.95	8.03
	Physical Education	866	43.61	8.80	99	41.55	10.38	767	43.88	8.55
	Secondary English-Language Arts	1879	46.89	6.97	410	45.56	6.56	1469	47.26	7.04
	Secondary History/Social Studies	1786	43.66	6.79	391	41.72	7.08	1395	44.20	6.61
	Secondary Mathematics	1547	42.63	6.92	307	40.91	7.34	1240	43.05	6.75
	Secondary Science	1248	45.18	7.63	220	43.42	7.97	1028	45.56	7.51
	Special Education	3043	40.89	8.28	678	39.67	8.00	2365	41.24	8.33
	Technology and Engineering Education	51	42.20	8.70	8	42.75	6.30	43	42.09	9.14
	Visual Arts	603	46.22	7.24	125	43.72	7.33	478	46.87	7.08
	All	27172	44.24	7.42	5720	43.15	7.53	21452	44.53	7.36

13 Rubrics	**Field**	N	Mean	Std. Deviation	N	Mean	Std. Deviation	N	Mean	Std. Deviation
	Classical Languages	15	31.73	6.62	2	30.50	7.78	13	31.92	6.76
	World Language	572	37.24	7.39	93	36.44	6.46	479	37.39	7.55
	All	587	37.10	7.42	95	36.32	6.49	492	37.25	7.58

18 Rubrics	Field	N	Mean	Std. Deviation	N	Mean	Std. Deviation	N	Mean	Std. Deviation
	Elementary Education	3869	53.67	9.09	457	52.00	9.40	3412	53.89	9.03
	All	3869	53.67	9.09	457	52.00	9.40	3412	53.89	9.03

Appendix G: ANOVAs and Post-hoc Analyses

One-way ANOVAs were run to examine significance of differences between subgroups in each demographic field. Post-hoc comparisons using the Games-Howell procedure, which does not rely on the assumption of equal variance between subgroups, were then considered to analyze differences within each category.

Note: Analyses presented do not include portfolios that do not fall into an interpretable category for that demographic field (i.e.: other, unidentified) or have a sample size of less than 100. Due to unequal sample sizes and variances between subgroups, all comparisons should be interpreted with caution.

Table 1: Teaching Placement Context

ANOVA

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	13237.849	4	3309.462	61.705	.000
Within Groups	1150278.825	21447	53.634		
Total	1163516.674	21451			

Post Hoc Analyses

(I) Teaching Context	(J) Teaching Context	Mean Difference (I-J)	Std. Error	Sig.
Rural	Rural/Suburban	-.057	.213	.999
	Suburban	-1.963*	.147	.000
	Suburban/Urban	-.886*	.198	.000
	Urban	-1.603*	.152	.000
Rural/Suburban	Rural	.057	.213	.999
	Suburban	-1.906*	.194	.000
	Suburban/Urban	-.828*	.236	.004
	Urban	-1.546*	.198	.000
Suburban	Rural	1.963*	.147	.000
	Rural/Suburban	1.906*	.194	.000
	Suburban/Urban	1.078*	.179	.000
	Urban	.360*	.125	.032
Suburban/Urban	Rural	.886*	.198	.000
	Rural/Suburban	.828*	.236	.004
	Suburban	-1.078*	.179	.000
	Urban	-.718*	.183	.001
Urban	Rural	1.603*	.152	.000
	Rural/Suburban	1.546*	.198	.000
	Suburban	-.360*	.125	.032
	Suburban/Urban	.718*	.183	.001

*. The mean difference is significant at the 0.05 level

Table 2: Ethnicity

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	15952.232	7	2278.890	42.585	.000
Within Groups	1147564.442	21444	53.514		
Total	1163516.674	21451			

Post Hoc Analyses

(I) Ethnicity	(J) Ethnicity	Mean Difference (I-J)	Std. Error	Sig.
African American/Black	American Indian or Alaskan Native	-2.579*	.789	.034
	Asian or Pacific Islander	-4.399*	.321	.000
	Hispanic	-3.625*	.300	.000
	White	-3.621*	.230	.000
Asian or Pacific Islander	African American/Black	4.399*	.321	.000
	American Indian or Alaskan Native	1.820	.791	.308
	Hispanic	.774	.306	.184
	White	.778*	.238	.025
Hispanic	African American/Black	3.625*	.300	.000
	American Indian or Alaskan Native	1.045	.783	.882
	Asian or Pacific Islander	-.774	.306	.184
	White	.004	.209	1.000
White	African American/Black	3.621*	.230	.000
	American Indian or Alaskan Native	1.042	.759	.866
	Asian or Pacific Islander	-.778*	.238	.025
	Hispanic	-.004	.209	1.000

Table 3: Primary Language

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	426.397	1	426.397	7.869	.005
Within Groups	1149068.283	21206	54.186		
Total	1149494.680	21207			

Table 4: Gender

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7946.588	1	7946.588	147.713	.000
Within Groups	1140130.113	21193	53.797		
Total	1148076.701	21194			

Table 5: Education level

ANOVA[a]					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6570.262	3	2190.087	40.60	<.0001
Within Groups	1156946.412	21448	53.942		
Total	1163516.674	21451			

(I) Race	(J) Race	Mean Difference (I-J)	Std. Error	Sig.
High school/Some college	Bachelor's/Bachelor's plus credits	-1.11211*		<.05
	Master's/Master's plus credits	-0.02653		
	Doctorate	-2.22918		

Appendix H: Demographic subgroups within teaching context

The following tables present cross-tabs breakdowns of candidates' ethnicity and gender within each teaching context.

Table 1: Ethnicity by Teaching Context

Teaching Context	Ethnicity	Mean	N	Std. Deviation
Rural	**African American/Black**	39.67	159	7.82
	American Indian or Alaskan Native	41.59	27	6.79
	Asian or Pacific Islander	44.38	42	5.68
	Hispanic	42.92	128	6.96
	White	43.22	4564	7.42
	Multiracial	43.38	64	6.79
	Other	41.83	24	7.72
	Undeclared	43.26	81	7.85
	All	43.10	5089	7.43
Rural/Suburban	**Ethnicity**			
	African American/Black	39.45	80	9.03
	American Indian or Alaskan Native	40.00	7	7.30
	Asian or Pacific Islander	41.74	34	7.36
	Hispanic	42.58	69	5.84
	White	43.44	2129	7.40
	Multiracial	43.08	48	7.08
	Other	41.41	17	6.41
	Undeclared	43.69	74	8.99
	All	43.24	2458	7.49

Suburban	**Ethnicity**			
	African American/Black	41.88	296	7.22
	American Indian or Alaskan Native	45.79	19	4.42
	Asian or Pacific Islander	46.01	302	6.74
	Hispanic	44.23	414	7.87
	White	45.01	7413	7.10
	Multiracial	44.83	207	6.97
	Other	45.59	80	8.49
	Undeclared	45.80	221	7.88
	All	44.93	8952	7.18
Suburban/Urban	**Ethnicity**			
	African American/Black	39.30	159	7.84
	American Indian or Alaskan Native	44.17	6	6.97
	Asian or Pacific Islander	44.09	121	7.60
	Hispanic	44.41	126	6.96
	White	44.14	2172	7.25
	Multiracial	45.30	99	7.11
	Other	43.83	23	7.95
	Undeclared	45.27	94	8.10
	All	43.95	2800	7.40

Urban	**Ethnicity**			
	African American/Black	41.51	740	7.65
	American Indian or Alaskan Native	42.32	19	8.29
	Asian or Pacific Islander	45.20	470	6.83
	Hispanic	45.09	745	7.36
	White	44.98	5164	7.42
	Multiracial	44.98	254	8.14
	Other	44.05	176	7.66
	Undeclared	43.83	305	8.39
	All	44.60	7873	7.54

Table 2: Gender by Teaching Context

Teaching Context	Gender	Mean	N	Std. Deviation
Rural	Male	41.86	1099	7.922
	Female	43.44	3945	7.259
	Total	43.10	5044	7.437
Rural/Suburban	Male	42.00	672	7.773
	Female	43.72	1747	7.315
	Total	43.24	2419	7.483
Suburban	Male	43.73	1851	7.603
	Female	45.24	7020	7.037
	Total	44.93	8871	7.185
Suburban/Urban	Male	42.43	668	7.652
	Female	44.45	2092	7.221
	Total	43.96	2760	7.377
Urban	Male	43.63	1707	7.963
	Female	44.87	6047	7.387
	Total	44.60	7754	7.535

Appendix I: National Technical Advisory Committee (TAC)

Members	Institution
Andrew Porter	University of Pennsylvania
Jim Pellegrino	University of Illinois at Chicago
Pam Moss	University of Michigan
Andy Ho	Harvard University
Lloyd Bond	Carnegie Foundation
Brian Gong	The National Center for the Improvement of Educational Assessment
Stuart Kahl	Measured Progress
Eva Baker	University of California, Los Angeles
Jamal Abedi	University of California, Davis
Edward Haertel	Stanford University
Mark Wilson	University of California, Berkeley
Lorrie Shepard	University of Colorado, Boulder
Linda Darling-Hammond	Stanford University
Ruth Chung Wei	Stanford University
David Pearson	University of California, Berkeley
Anthony S. Bryk	The Carnegie Foundation
Susanna Loeb	Stanford University
James Popham	University of California, Los Angeles

All members of the national technical advisory committee were presented with the draft version of this report and had an opportunity to provide comments and feedback. Several other experts in the field were consulted and provided valuable recommendations. We thank them for their ongoing input and guidance.

Citations

Adkins, A., Klass, P., & Palmer, E. (2015). Identifying demographic and preserve teacher performance predictors of success on the edTPA. Paper Presented at the 2015 Hawaii International Conference on Education. Honolulu, Hawaii.

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.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. Educational Assessment, 17(2-3), 62–87. doi:10.1080/10627197.2012.715014

Benner, S.M. & Wishart, B. (2015) Teacher preparation program impact on student learning: Correlations between edTPA, and VAM levels of effectiveness. Paper presented at the 2015 annual meeting of the meeting of the American Educational Research Association, Chicago, IL.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. Phi Delta Kappan, 80(2), 139–148.

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.

Cavalluzzo, L., Barrow, L., Mokher, C., Geraghty, T., & Sartain, L. (2014). From Large Urban to Small Rural Schools: An empirical study of National Board certification and teaching effectiveness. Alexandria, VA: The CNA Corporation. Retrieved from http://www.cna.org/sites/default/files/research/IRM-2015-U 010313.pdf.

Cowan, J., & Goldhaber, D. (2015). National Board Certification and Teacher Effectiveness: Evidence from Washington. Technical Report 2015-1, Center for Education Data and Research, Seattle, WA. Retrieved from http://www.cedr.us/papers/working/CEDR%20WP%2020153_NBPTS%20Cert.pdf.

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

Chung, R. R. (2008). Beyond assessment: Performance assessments in teacher education. Teacher Education Quarterly, 35(1), 8-28.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

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

Darling-Hammond, L. (2010). Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching. Washington, DC: Center for American Progress.

Darling-Hammond, L., & Falk, B. (2013). Teacher learning through assessment: How student-performance assessments can support teacher learning. Center for American Progress. Retrieved from www.americanprogress.org.

Darling-Hammond, L., Newton, S. P., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. Educational Assessment, Evaluation and Accountability, 25(3), 179-204.

Duckor, B., Castellano, K. E., Tellez, K., Wihardini, D., & Wilson, M. (2014). Examining the internal structure evidence for the Performance Assessment for California Teachers: A validation study of the Elementary Literacy Teaching Event for Tier I teacher licensure. Journal of Teacher Education, 65(5), 402–420. http://doi.org/10.1177/0022487114542517

Gardner, P. L. (1970), Test Length and the Standard Error of Measurement. *Journal of Educational Measurement*, 7: 271–273.

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.

Gillham, J.C. & Gallagher, D. (2015). Pilot Implementation of the edTPA in Ohio. Paper presented at the 2015 annual meeting of the American Association of Colleges for Teacher Education, Atlanta, GA.

Haertel, E. H. (2008). Standard setting. In K. E. Ryan & L. A. Shepard (Eds.), The future of test-based educational accountability (pp. 139–154). New York: Taylor & Francis.

Haertel, E. H., Beimers, J. N., & Miles, J. A. (2012). The briefing book method. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed., pp. 283–299). New York, NY: Routledge.

Haertel, E. H., & Lorié, W. A. (2004). Validating standards-based test score interpretations. Measurement: Interdisciplinary.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. Measurement: Interdisciplinary Research & Perspective, 2(3), 135–170.

Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th ed., pp.17–64). Westport, CT: American Council on Education and Praeger.

Kiefer, T., Robitzsch, A., & Wu, M. L. (2015). Test analysis modules (Version 1.6-0) [R Package]. Retrieved from http://cran.us.r-project.org/web/packages/TAM/index.html

Kleyn, T., López, D., & Makar, C. (2015). What About Bilingualism? A Critical Reflection on the edTPA With Teachers of Emergent Bilinguals, Bilingual Research Journal: The Journal of the National Association for Bilingual Education, 38:1, 88-106, DOI: 10.1080/15235882.2015.1017029

Lin, S. (2015) Learning through Action: Teacher Candidates and Performance Assessments. Doctoral dissertation, University of Washington, Seattle, WA.

Liu, L. B., & Milman, N. B. (2013). Year one implications of a teacher performance assessment's impact on Multicultural Education across a secondary education teacher preparation program. Action in Teacher Education, 35(2), 125-142.

Lord, F. M. (1959). Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 19, 233-239.

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

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. Improving Schools, 12(2), 101-113.

Otero, V. K. (2006). Moving beyond the “Get it or don’t” conception of formative assessment. Journal of Teacher Education, 57, 247–255.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). Journal of Teacher Education, 57(1), 22–36. doi:10.1177/0022487105284045

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

Peck, C., Gallucci, C., & Sloan, T. (2010). Negotiating implementation of high-stakes performance assessment policies in teacher education: From compliance to inquiry. Journal of Teacher Education, 61(5), 451-463. doi: 10.1177/0022487109354520

Peck, C., Gallucci, C., Sloan, T., & Lippincott, A. (2009). Organizational learning and program renewal in teacher education: A socio-cultural theory of learning, innovation and change. Educational Research Review, 4, 16-25. doi:10.1016/j.edurev.2008.06.001

Peck, C., & McDonald, M. (2013). Creating "cultures of evidence" in teacher education: Context, policy and practice in three high data use programs. The New Educator, 9(1), 12-28. doi: 10.1080/1547688X.2013.751312

Peck, C. A., Singer-Gabella, M., Sloan, T., & Lin, S. (2014). Driving blind: Why we need standardized performance assessment in teacher education. Journal of Curriculum and Instruction, 8(1), 8-30.

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.

Sandholtz, J. H., & Shea, L. M. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. Journal of Teacher Education, 63(1), 39–50. doi:10.1177/0022487111421175

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? Journal of Teacher Education, 0022487114542518.

Shepard, L. A. (1993). Evaluating test validity. Review of research in education, 405-450.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. Journal of Science Teacher Education, 22(4), 371-391.

Sloan, T. (2013). Distributed leadership and organizational change: Implementation of a teaching performance measure. The New Educator, 9, 29-53. doi: 10.1080/1547688X.2013.751313

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

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. Best practices in quantitative methods, 29-4

Stillman, J., Anderson, L., Arellano, A., Lindquist Wong, P., Berta-Avila, M., Alfaro, C., & Struthers, K. (2013). Putting PACT in context and context in PACT: Teacher educators collaborating around program-specific and shared learning goals. *Teacher Education Quarterly*, 40(4), 135-157.

Tennessee Department of Education (2016). Subcommittee on Educator Preparation and Licensing, State Board of Education, May 19 and August 12, 2016, Nashville, TN.

Whittaker, A., & Nelson, C. (2013). Assessment with an “End in View”. *The New Educator*, *9*(1), 77-93.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modeling software*. ACER Press.