Affirming the Validity and Reliability of edTPA

A response authored by the Stanford Center for Assessment, Learning, and Equity (SCALE) and Pearson

December 16, 2019

The article "Assessing the Assessment: Evidence of Reliability and Validity in the edTPA" (Gitomer, Martinez, Battey & Hyland, 2019) raises questions about the technical documentation and scoring of edTPA. We respond to these questions by providing detailed information about edTPA's development as a subject-specific assessment with a shared and common pedagogical architecture; clarification on technical documentation; and corrections on the inaccurate representation of edTPA's scoring model, the implementation of double scoring, and the operational safeguards in place to monitor scoring accuracy. This response also addresses the many inaccuracies and misunderstandings about edTPA within the article.

edTPA was collaboratively designed by teacher educators and P–12 teachers nationwide under the leadership of the Stanford Center for Assessment, Learning and Equity (SCALE). In its seven years of operation, edTPA has accumulated and publicly released a substantial and growing body of technical evidence supporting its intended use as a measure of readiness to teach, and as a means to inform program approval or accreditation. This evidence demonstrates that edTPA is a valid and reliable support and assessment system appropriate for use as a high-stakes assessment.

We categorically reject the conclusions of the recent Gitomer et. al. (2019) article and firmly reject the call for a moratorium on the use of edTPA.

Section I – Purpose and Design of edTPA

edTPA was designed as an educative teacher performance support and assessment system based on three foundational pillars that guided its original development: (1) ground-breaking theory based on research by Lee Shulman that describes the role of Pedagogical Content Knowledge (PCK) in preparing and assessing teachers (Shulman, 1986); (2) the design and development of the National Board for Professional Teaching Standards' (NBPTS) assessment of accomplished teaching, which was built from Shulman's early research[1] as well as the principles of the Interstate Teacher Assessment and Support Consortium (InTASC) standards; and (3) the definition of edTPA as *educative*, which represents a commitment and responsibility to build a support and assessment system "of and for learning".

An Authentic Representation of Teaching. 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 (SCALE, 2015) provides the conceptual and empirical rationale for edTPA's approach to assessing content pedagogical knowledge through an authentic cycle of teaching represented in a three-task design (i.e., planning, instruction, and assessment) focused on student learning. That research base also informs the representation in the scoring rubrics of initial competencies needed to be ready to teach.

Authentic evidence found across the three tasks includes lesson plans, instructional materials, student assignments and assessments, feedback on student work, and unedited video of instruction from the

[1] edTPA author, Dr. Ray Pecheone, was a member of the team that designed the portfolio system adopted as the NBPTS framework.

candidate's internship or placement. Also assessed through the three tasks are candidates' abilities to develop their students' academic language and justify and analyze their teaching practices. Candidates' commentaries reveal their decision making in relation to their knowledge of students and subject matter, as well as reflections on how their teaching decisions and strategies have influenced student learning. edTPA evaluates candidates on a five-point scale across 15 rubrics.

As Gitomer and colleagues (2019) acknowledge, "The literature and available edTPA documentation provide considerable support for the claim that the assessments capture important aspects of quality teaching that should be considered in teacher certification." This view of edTPA is shared by the hundreds of teachers and teacher educators who participated in the development and review of the assessment over the several years in which it was built and refined, and by many more who continue to use its results to guide ongoing decisions about program improvement. Teachers and teacher educators find it both measures and helps develop teaching knowledge and skill.[1]

Common Architecture. The 28 edTPA subject-specific assessments and their associated handbooks share approximately 80% of their design, assessing common 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. The structure of the 28 handbooks includes a common architecture, common pedagogical language, common set of identical prompts and rubrics to document a prospective teacher's practice. Although Gitomer et. al. (2019) suggest that aggregate analyses (statistics that examine all subject areas in combination) are inappropriate, this common architecture makes such analyses reasonable and sensible.

Transparency and Technical Documentation. Performance data, reliability, and validity evidence are publicly documented by SCALE in the annual edTPA administrative reports that support states and programs in the review and use of edTPA. In 2013 SCALE released the Field Test Summary Report (SCALE, 2013) to document the design, development process, and validation results from multi-year pilot and field-testing efforts. The initial edTPA administrative report was released in 2015 (SCALE, 2015) and has been produced annually to describe the operational administrations from the prior calendar year. We believe edTPA publishes as much as, if not more, data each year on its assessment, scoring, and other important edTPA program aspects than most other similar assessments.

Effective Use. Designed as a capstone assessment, edTPA is a meaningful component within a support and assessment system used in Educator Preparation Programs (EPPs)as one measure of teacher readiness. SCALE's commitment to supporting faculty and teacher candidates in this educative process has resulted in supports such as (1) the production and publication of nearly 200 faculty and SCALE-developed resources supporting the implementation of edTPA[2]; (2) the development of a national academy of faculty offering training and workshops to EPPs; and (3) the support and sponsorship of both annual and regional conferences to enable participating faculty to share research and practice.

The design is appropriate for high-stakes decision making because it has been validated through multiple means, field tested, and implemented with a high degree of quality control and accuracy.

Section II –Validity Evidence

While purposes and effects of the assessment are very important, technical reliability and validity are equally important, and discussed in the following sections. The edTPA program demonstrates technical reliability and validity with a high level of quality to meet the policy and implementation needs of State Boards of Education, Independent Standards Boards, and State Education Agencies. The edTPA also addresses the needs of EPPs using the assessment to inform decisions about curriculum improvement.

In educational measurement and assessment, validity claims are connected to the soundness of interpreting decisions and actions: "A validity theory provides guidance about what it means to say that an interpretation, decision, or action is more or less sound; about the sorts of evidence, reasoning, and criteria by which soundness might be judged; and about how to develop more sound interpretations, decisions, and actions" (Moss, Girard, & Haniford, 2006).

Validity evidence is used to establish that the assessment properly measures the constructs or concepts it intends to measure (construct validity) and that it measures content that is important for drawing the inferences the test is meant to inform (content validity). For licensure tests, content validity is essential, but perhaps the most compelling validity evidence is that associated with predictive validity: That is, does the assessment actually predict teachers' effectiveness in the classroom? edTPA has and continues to develop a strong body of validity evidence.

Construct Validity. edTPA is widely used as a performance-based assessment for evaluating readiness to teach because it measures what is important to teaching. SCALE conducted a review of over 200 relevant studies (released as a research synthesis) and used this review to inform the design principles of, and foundation for, the common architecture of edTPA (SCALE, 2015). This research, along with the related standards of the NBPTS and InTASC, informs the representation of effective teaching and the underlying constructs evaluated by the 15 rubrics used in edTPA. A research study that evaluated the theoretical constructs underlying the development of edTPA concluded that "content and construct validity can be argued as being technically sound in the edTPA—when these arguments are also based on other professionalizing efforts such as the establishment of InTASC teaching standards and the performance assessment model of the NBPTS" (Sato, 2014).

Subject-Specific Content Validity. Gitomer, et. al (2019) question the content validation of the edTPA, suggesting "there is no information available indicating that the specific and detailed content analysis for each of the 27 assessments have been conducted." This assertion is not correct. edTPA has established content validity through several means.

First, subject-specific, expert design teams provided content validity evidence of the specific job-related competencies assessed within each subject area and guided the development of edTPA handbooks across 28 subject area fields. edTPA design teams included P–12 teachers, teacher-leaders, and faculty with deep subject-matter expertise, strong subject-matter pedagogy, and extensive experience working with pre-service teachers. The design teams participated in a process using subject-matter-specific content and pedagogical standards to determine the types of teaching and learning that edTPA handbooks would emphasize for their field.

Second, a separate job analysis study was conducted with an additional pool of educators who provided evidence to confirm the degree to which the job requirements of a teacher are aligned to edTPA. Survey design and data analysis for the job analysis were conducted by the Human Resources Research Organization (HumRRO). The survey data reinforced the strength of the relationship between edTPA tasks and rubrics and the critical tasks performed by educators (HumRRO, 2013).

Third, a separate content validation committee was convened to independently provide ratings on the importance, alignment, and representativeness of the knowledge and skills required for each edTPA rubric, in relation to national pedagogical and content-specific standards (SCALE, 2013).

Predictive Validity. Teacher licensure assessments have rarely established evidence that their outcomes are actually related to candidates' performance on the job. However, studies involving the Performance

Assessment for California Teachers (PACT), the predecessor of edTPA, revealed predictive validity evidence in mathematics achievement and literacy achievement (Darling-Hammond, Newton & Chung, 2013). In addition, Goldhaber, Cowan and Thoebald (2016) found that "passing the edTPA is significantly predictive of teacher effectiveness in reading." Additional research detailing validity analyses of first-year teachers' value-added estimates and evaluation ratings shows that edTPA outcomes significantly predict first-year teacher performance (see Bastian, Henry, Pan and Lys, 2016).

Section III – Reliability of Scoring

The edTPA portfolio is a collection of performance-based activities that requires human scoring. Because human judgments naturally contain more variation or measurement error than more simplified multiple-choice test questions that are machine scored, accuracy of scoring is shaped through a comprehensive set of quality control scoring and management practices. While traditional quantifications of such variation are used for edTPA (such as kappa-n), these statistics are used in conjunction with rigorous quality control procedures employed to help ensure consistent and accurate scoring.

It is critical in licensure assessments to ensure reliability around the cut score. Gitomer et. al (2019) raise concerns about the scoring processes of edTPA, which they appear to misunderstand. The management of scoring to produce consistency and careful judgments, especially around the cut score, is detailed below.

There are several steps that aim to support a consistency in classification, including an intentional scoring design appropriate to measure the edTPA cycles of teaching, a strategic double-scoring plan, and stringent methods for selection, qualification, calibration, training and monitoring. Guidance from the edTPA National Technical Advisory Committee informed the scoring model and quality management plan for the development and launch of edTPA. The results of edTPA scorer reliability 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 and valid estimate of a prospective teacher's readiness to teach.

Intentional Scoring Design. The design of edTPA includes the goal of evaluating the "totality of the evidence" submitted in each portfolio in order to understand how the teacher engages in a cycle of teaching. A scorer views the entire teaching cycle, so that the integrated teaching and learning experiences are judged in totality, not in separate parts. Having the same scorer evaluate all three tasks of the assessment (planning, instruction, and assessment) allows the scorer to obtain a comprehensive picture of the candidate's performance, understanding how decisions made in the context of planning around student needs are also relevant to subject-specific instruction and the assessment of student learning, for example (Pecheone & Chung, 2006).

Subject Specific Scoring. The measurement of each portfolio's content is enhanced by the subject-specificity embedded in the scoring model. edTPA assessments are evaluated only by qualified scorers based on their content knowledge, pedagogical and grade-level expertise, and experience working with pre-service teachers and only so after rigorous training and qualification. Scorers are specifically trained by content area and score only assessments in the content areas and grade spans for which they have verified expertise and qualifications. This enhances the likelihood that the validity of the content elicited by the prompts is properly recognized in the scoring process.

Double Scoring Model. Gitomer et. al (2019) appear to argue that all portfolios should be double scored, and state "standard practice involves collecting multiple independent measures or instances of practice". These arguments are inaccurate as the current industry standard for portfolio-based teacher credentialing assessments does not apply a 100% double-scoring model. For example, the National Board Certification

exams, ETS's Praxis Performance Assessment for Teachers (PPAT), and the ProTeach exam are not 100% double scored but rather use models similar to edTPA to selectively double score, and to supplement quality management steps.[3, 4, 5] Examinations in other professions, such as the Bar exam's Multistate Essay Examination (MEE) and the Multistate Performance Test (MPT), as well as the Certified Professional Accounts (CPA) exam, also do not use 100% double scoring.

The statement from Gitomer et. al (2019) about the amount of double scoring that occurs in edTPA is inaccurate. More than 30% of edTPA portfolios are double scored, including an independent 10% random selection and over 20% of portfolios that fall within the double-scoring band as described in the scoring model below.

Resolution and Adjudication. In addition to 10% random double scoring, portfolios within a defined range above and below the state-specific cut score (or national cut score, if there is no state-specific cut score in place) are scored by two, and sometimes three, scorers. In all such cases the final score is based on at least two scorers who agree. Adjudication and resolution are also implemented as part of double scoring. If Scorer 1 and Scorer 2 are more than 1 score point apart on any rubric, the portfolio is resolved by a scoring supervisor. If total scores for all 15 rubrics from Scorer 1 and Scorer 2 are on opposite sides of the cut score, the portfolio is adjudicated by a scoring supervisor. If scorers have five or more adjacent rubric scores (scores that are one point apart) out of 15 rubrics, the portfolio is also scored by a scoring supervisor. These double-scoring procedures support the final scores assigned to edTPA candidates.

Scoring Quality Management. Scorer quality management in edTPA is designed for consistency and accuracy in scoring. That includes extensive scorer training and oversight, rigorous calibration procedures, the selection of expert scorers who meet strict qualifications, scorer drift monitoring, and a score appeal process.

Scoring for edTPA occurs year-round (as contrasted with other portfolio assessments that utilize event-based scoring sessions) and therefore scorer quality monitoring is done on a constant basis. edTPA implements multiple quality management measures to help ensure reliable scoring, including the use of benchmark portfolios, systematic backreading for each scorer, and a scoring model that addresses scoring accuracy through comprehensive resolution and adjudication around the cut scores.

Benchmarks. Scorers are periodically assigned benchmark assessments as calibration exercises providing valuable information about a scorer's performance in relation to an established set of scores. Double scoring is utilized to monitor agreement rates amongst scorers, providing inter-rater reliability metrics.

Backreading. As is common in human scoring, backreading—defined as supervisors scoring a previously scored portfolio for the purpose of reviewing the original scores and providing feedback to the scorer—is also utilized to monitor ongoing scorer performance. Newly qualified scorers are each backread by a scoring supervisor after they have scored their first assessment and prior to the release of those scores. All scorers are systematically monitored via the backreading process to help ensure their scores are accurate and reliable.

Scorer Agreement Statistics. The edTPA technical reports include evidence regarding consistency in scoring. One component of that report examines inter-rater consistency, which we evaluate using the kappa-n statistic. Gitomer et. al (2019) criticize the use of this statistic and further suggest that "edTPA uses its own version of Kappa." This is incorrect. The kappa-n statistic is widely used in the field and is appropriate to the context of edTPA.

Research shows that for scorers who assign scores to categories at random, but in line with expected base rates, kappa-n should accurately reflect agreement attributable to scorers (Hsu & Field, 2003). Given the

training and qualification of edTPA scorers, the expectation is that there will be large base rate agreement and, therefore kappa-n was chosen as the agreement statistic to use. While Gitomer et. al (2019) argue that the kappa-n statistic can sometimes appear inflated, this is only the case when scorers radically disagree about base rates, a theoretical condition that does not occur operationally given the training and qualification of edTPA scorers. Statistics such as Kappa (including kappa-n) are intended to be used with other methods of consistency as outlined. Therefore, the body of evidence provided needs to be considered and supplements the interpretation of any one statistic.

The edTPA authors have defined total agreement as being represented by adjacent and perfect agreement and applied that definition in the kappa analysis. Given the three possible classifications of agreement (perfect, adjacent, and non-agreement) in quantifying how much agreement there is, perfect and adjacent were combined as the agreement statistic (keeping in mind scoring rules allowing only up to five adjacent scores out of 15). Due to the design as an integrated cycle of teaching, 15 rubrics each on a five-point scale, with an accumulated total score, agreement at exact and adjacent is a reasonable statistical representation of total agreement.

Similarly, regarding the estimation of the Standard Error of Measurement, Gitomer, et. al. (2019) indicate that the formula was nonstandard and that it cannot be found in the literature. In fact, the formula used for edTPA is Lord's 'Method Three' formula (1984, pg. 241) which Lord himself calls an 'unbiased estimate' that holds 'without further assumptions'.

Gitomer and colleagues (2019) pose questions about the use of the same cut score for the assessment across subjects. edTPA measures what is important in teaching and is intentionally designed as a performance assessment that evaluates a common set of teaching principles and teaching behaviors. edTPA uses rubrics that describe teaching competencies that tap into this deep and common structure while also allowing for subject specific variances. This rationale behind edTPA was supported via the extensive job analyses conducted, demonstrating that such knowledge and skills are supported across fields. As such, the concept of a single passing standard rooted in this common architecture and integration of the three task, fifteen rubric structure of edTPA is appropriate. edTPA's rubrics capture performance common across disciplines as well as performance to specific disciplines. This is also observed in other portfolio measures of teaching such as NBPTS.

Based on these commonalities, edTPA's standard setting methodology uses a single, shared standard regarding readiness to teach can be achieved regardless of subject area. edTPA intentionally chooses to treat candidates who earn equivalent total scores as demonstrating equivalent readiness to teach. This acknowledges that candidates may be stronger in one dimension than another, but if their total performance reaches a given threshold, they will be considered to have demonstrated a sufficient level of performance.

Section IV – Conclusion

The edTPA program has a demonstrated commitment to quality, continual improvement, and the collection of additional reliability and validity information. Plans remain in place for studies which will be publicly released, for example, a generalizability analysis as well as additional interscorer agreement statistics. Furthermore, the annual edTPA administrative report will continue to add more evidence of edTPA's appropriate use for licensure and certification.

edTPA is a widely used and rigorously field-tested performance assessment of new teachers that meets industry standards as represented by the technical standards of AERA, APA, & NCME (2014). Evidence that the assessment is valid and reliable is publicly reported annually in its administrative reports as well as in a number of other technical resources available at the online Resource Library at edtpa.aacte.org. The substantial and ongoing collection of evidence of edTPA's technical qualities, based on its nationwide use

over the past several years, firmly supports its intended use to support licensing decisions. We categorically reject the conclusions of the recent AERJ article and firmly reject the call for a moratorium on the use of edTPA.

References

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

Bastian, K., Henry, G., Pan, Y., & Lys, D. (2016) "Teacher candidate performance assessments: Local scoring for teacher preparation program improvement." Teaching and Teacher Education, Volume 59, October 2016, pages 1-12.

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.

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.

Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2019). Assessing the Assessment: Evidence of Reliability and Validity in the edTPA. American Educational Research Journal. https://doi.org/10.3102/0002831219890608

Goldhaber, D., Cowan, J., and Theobald, R. (2016). Evaluating Prospective Teachers: Testing the Predictive Validity of the edTPA. CEDR Working Paper 2016-2.2. University of Washington, Seattle, WA. http://www.cedr.us/papers/working/CEDR%20WP%202016-2.2.pdf

Hsu, L. M., & Field, R., (2003). Interrater Agreement Measures: Comments on Kappan, Cohen's Kappa, Scott's π, and Aickin' s α.. Understanding Statistics, 2 (3), 205-219.

HumRRO (2013). National Teacher Job Analysis: Task Analysis Questionnaire (TAQ). Prepared for the edTPA program, 2013.

Lord, M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, 21(3), 239- 243.

Moss, P., Girard, B., & Haniford, L. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162. doi: 10.3102/0091732X030001109

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

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education,* 65, 421-424 doi:10.3102100283120831695 5

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. Retrieved from https://pdfs.semanticscholar.org/f29d/a5d8c806102b060e7669f67b5f9a55d8f7c4.pdf

Stanford Center for Assessment, Learning and Equity (SCALE). (2013). edTPA Field test: Summary report. Palo Alto, CA: Author. Retrieved from https://secure.aacte.org/apps/rl/res_get.php?fid=827&ref=edtpa

Stanford Center for Assessment, Learning and Equity (SCALE). (2015). Review of research on teacher education edTPA task dimensions and rubric constructs. Palo Alto, CA: Author.

https://scale.stanford.edu/sites/default/files/edTPA%20Literature%20Review%20Version2%20FINAL.pdf

Stanford Center for Assessment, Learning and Equity (SCALE). (2019). *Educative assessment and meaningful support: 2018 edTPA administrative report*. Palo Alto, CA: Author. Retrieved from: https://secure.aacte.org/apps/rl/res_get.php?fid=4769&ref=edtpa

Wilson, M., Hallam, P. J., Pecheone, R. L., Moss, P. A. (2014). Evaluating the Validity of Portfolio Assessments for Licensure Decisions. *Education Policy Analysis Archives, 22*(6). Retrieved from http://dx.doi.org/10.14507/epaa.v22n6.2014

---

[1] See: http://edtpa.aacte.org/voices-from-the-field
[2] See: https://secure.aacte.org/apps/rl/resource.php?ref=edtpa
[3] http://www.nbpts.org/wp-content/uploads/NBPTS_Scoring_Guide.pdf, p.8
[4] https://www.ets.org/ppa/test-takers/teachers/scores/how/
[5] http://www.waproteach.com/scoring/how_scored.html