Who’s assessing the assessment? The cautionary tale of the edTPA

Drew H. Gitomer, Jose Felipe Martinez, and Dan Battey
February 22, 2021

What will it take to hold test developers accountable for their practices?

The use of high-stakes assessments in public education has always been contested terrain. Long-simmering debates have focused on their benefits, the harms they cause, and the roles they play in decisions about high school graduation, school funding, teacher certification, and promotion. However, for all the disagreement about how such assessments affect students and teachers, and how they should or should not be used, it has generally been assumed that the assessment instruments *themselves* follow standard principles of measurement practice.

At the most basic level, test developers are expected to report truthful and technically accurate information about the measurement characteristics of their assessments, and they are expected to make no claims about those assessments for which they have no supporting evidence. Violating these fundamental principles compromises the validity of the entire enterprise. If we cannot trust the quality of the assessments themselves, then debates about how best to use them are beside the point.

Our research suggests that when it comes to the edTPA (a tool used across much of the United States to make high-stakes decisions about teacher licensure), the fundamental principles and norms of educational assessment have been violated. Further, we have discovered gaps in the guardrails that are meant to protect against such violations, leaving public agencies and advisory groups ill-equipped to deal with them. This cautionary tale reminds us that systems cannot counter negligence or bad faith if those in position to provide a counterweight are unable or unwilling to do so.

Background: Violations of assessment principles

The edTPA is a system of standardized portfolio assessments of teaching performance that, at the time this research was conducted, was mandated for use by educator preparation programs in 18 states, and approved in 21 others, as part of initial certification for preservice teachers. It builds on a large body of research over several decades focused on defining effective teaching and designing performance assessments to measure it. The assessments were created and are owned by Stanford Center for Assessment, Learning, and Equity (SCALE) and are now managed by Pearson Assessment, with endorsement and support from the American Association of Colleges for Teacher Education (AACTE). By 2018, just five years after they were introduced, they were among the most widely used tools for evaluating teacher candidates in the United States, reaching tens of thousands of candidates in hundreds of programs across the country. They have substantially influenced programs of study in teacher education. And for the teaching candidates who take them, they are a major undertaking, requiring them to make a substantial time investment, as well as costing them $300.

Getty Images

In 2018, two of us (Drew Gitomer and José Felipe Martínez) participated in a symposium at the annual meeting of the National Council of Measurement in Education (NCME), which included a presentation on edTPA by representatives of Pearson and SCALE (Pecheone et al., 2018). We were struck by specific claims that were made in that presentation: Reported rates of reliability seemed implausibly high, and reported rates of rater error seemed implausibly low, implying that a teaching candidate would receive the same scores regardless of who rated the assessment. A well-established feature of performance measures of teaching, similar to those being used in edTPA, is that raters will often disagree on their scores of any single performance and, therefore, the scoring reliability of any single performance is inevitably quite modest. The raw data on rater agreement that edTPA reports are consistent with the full body of work on these assessments. Yet, the reliabilities they reported, which depend on these agreement levels, were completely discrepant from all other past research.

At the NCME session, we publicly raised these concerns, and we offered to engage in further conversation to clarify matters and address our questions about the claims that were made. Upon further investigation, we found that the information presented at the session was also reported in edTPA’s annual technical reports — the very information state departments of education rely on to decide whether to use the edTPA for teacher licensure.

In December 2019, we published an article detailing serious concerns about the technical quality of the edTPA in the *American Educational Research Journal* (*AERJ*), one of the most highly rated and respected journals in the field of educational research (Gitomer et al., 2019). We argued that edTPA was using procedures and statistics that were, at best, woefully inappropriate and, at worst, fabricated to convey the misleading impression that its scores are more reliable and precise than they truly are. Our analysis showed why those claims were unwarranted, and we ultimately suggested that the concerns were so serious that they warranted a moratorium on using edTPA scores for high-stakes decisions about teacher licensure.

After the article was accepted for publication, we uncovered another apparent violation of basic norms of truthfulness. The edTPA technical reports released each year state that, “all analyses and results have been informed and reviewed by a technical advisory committee [TAC] of nationally recognized psychometricians” (Stanford Center for Assessment, Learning, and Equity, 2018, p. 1). The individuals listed are some of the most outstanding experts in the field of measurement, so, out of respect, we wrote to inform them about our forthcoming article. A majority of these experts responded that they were *not aware* of being listed as members of such a committee. Several recalled that they had participated in one or two initial conference calls, but they added that they had not been in touch with the program or reviewed any materials since its large-scale implementation in 2014. This direct quote represents a typical response:

> I think I can only recall 1 or 2 meetings that we had early on when we raised questions about samples, etc. But what is surprising is the statement that implies that the TAC signed off on the assessment as meeting the AERA, APA, NCME technical standards for licensure. I know of no such endorsement ever given by me as a member of the TAC.

Criticism and responses

Presenting such a critique and calling for a moratorium in an academic journal like *AERJ* is unusual, but we sought to publish the work in this manner for a few reasons. First, we wanted the study to go through a rigorous peer review process. Second, we wanted to avoid perceptions that we oppose testing in general or teacher testing in particular (we do not), so we did not consider outlets with well-defined agendas or political perspectives related to those topics. Third, we wanted to publish in a journal that would be widely read. Accordingly, the article was written in an academic register, using measured language. Yet, there could be no doubt that we were calling out malfeasance and highlighting a violation of principles that are the bedrock of any legitimate discussion of assessment validity. The paper received significant attention for an academic journal article and was downloaded substantially more than any other *AERJ* article during the six months after it was first published online.

While we thought a self-imposed moratorium by Pearson and SCALE was unrealistic, we did expect that others in the education sector — especially AACTE and individual state departments of education, which have oversight responsibilities related to prospective teachers, teacher education institutions, and the field of teaching more broadly — would try to understand the implications of our argument and take appropriate action. At a minimum, we expected these institutions to seek additional information to determine whether the claims in our article were true — either by asking independent experts to review our claims or by asking us to make our case to their respective boards or leadership groups. To our surprise, nothing of the sort happened.

This lack of public oversight and accountability has implications that extend far beyond the use of the edTPA itself.

Share this on ○ ○ ○

It is troubling that an assessment of this consequence, one that affects the lives of thousands of prospective teachers every year, does not appear to meet the minimum standards of the field of educational assessment. But it is even more concerning that there is presently no system in place to adjudicate and address a case like this, and the safeguards that do exist are woefully inadequate. The problems with edTPA most directly affect the preservice teachers who take this assessment and the organizations that use the scores. However, this lack of public oversight and accountability has implications that extend far beyond the use of the edTPA itself. Although we cannot ascribe motives or rationales to those involved, we believe that the response (and lack of response) to our concerns reveals serious problems not only with current approaches to teacher licensure but with the larger field of educational measurement and assessment.

Response from Pearson and SCALE

On December 5, 2019 (the day our article appeared online), Pearson and SCALE sent an email to each of their state contacts saying only that they stood by the technical quality of the edTPA and that "Technical experts at SCALE and Pearson are closely reviewing the *AERJ* article and will provide information soon." Indeed, on December 16, 2019, they released an eight-page response, titled *Affirming the Validity and Reliability of edTPA*, "categorically rejecting" our conclusions. The response, which was widely distributed to state departments of education and others in the broader edTPA network, framed our concerns either as resulting from our own inaccuracies and misunderstandings or as simple matters of reasonable professional disagreement, rather than as the more fundamental issues of potential malpractice we had claimed.

It is possible that by minimizing our concerns and describing them as falling within the bounds of normal professional debate, Pearson and SCALE persuaded AACTE and state departments of education to forego any further action. However, if this kind of unvetted corporate response had the effect of suppressing institutional reaction, then that would be highly problematic.
Had the edTPA organizations been able to mount a convincing rebuttal to the specific points made in our article, they should (and likely would) have submitted a response to the journal itself, subject to the same peer review process as our initial criticism. However, no such response has been submitted to *AERJ*, even though the journal's editors indicated they would welcome one. To date, apart from the corporate response by the edTPA organizations, there has not been a single substantive criticism of our paper, as far as we are aware.

Nevertheless, their unvetted response appears to have been effective in allaying state concerns. For example, a review by a working group adopted by the Connecticut State Board of Education in January 2020 acknowledged the *AERJ* study but accepted the unvetted Pearson/SCALE claims uncritically:

> Measures of reliability of edTPA are consistent with the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014).
>
> edTPA has consistently met or exceeded these standards for licensure. However, Gitomer, et al, 2019 recently called into question the reliability of the scoring of edTPA. SCALE responded by with [sic] a description of methods utilized in establishing reliability of scoring to address any misunderstandings about the scoring processes of edTPA. (McWeeney, 2020, p. 13)

Pearson and SCALE also raised questions about the motivations that led us to write the article, specifically claiming that the lead author (Drew Gitomer) had a conflict of interest because he previously had been an employee of Educational Testing Service (ETS), a competitor of Pearson in the teacher assessment field. We are aware of multiple exchanges, supported by minutes of state department of education meetings with Pearson, as well as confidential emails and conversations with multiple state department of education officials, all indicating that edTPA effectively used this argument to dismiss the concerns raised in our article. For example, working notes produced following a January 2020 meeting of the Professional Standard Board of the state of Washington with Ray Pecheone of SCALE state, "The author had a conflict of interest they did not disclose." Pearson also sent a written request that *AERJ* issue a correction or disclaimer noting the supposed conflict of interest. However, the request was rejected because there was no legitimate justification for any kind of conflict-of-interest disclosure. The employment in question, and any financial relationship between Gitomer and ETS, had ended eight years prior to the publication of the 2019 article — no statute defines employment conflicts as lasting in perpetuity, nor is it reasonable to assume that an author remains beholden to an organization he worked for many years ago.

Response from AACTE

The American Association of Colleges of Teacher Education (AACTE) is the leading teacher education association in the country, and it partners with SCALE and Pearson in endorsing edTPA. AACTE sees the assessment as a means of holding teacher education programs accountable for their performance, but doing so in a way that creates incentives for those programs to focus on high-quality teaching practices, using an assessment developed by members of the teacher education community (see https://edtpa.aacte.org/about-edtpa). However, early enthusiasm and broad support for the assessment has waned, as researchers and teacher education programs have raised significant concerns about the high-stakes version of edTPA and its negative impact on teachers and programs (e.g., Greenblatt, 2016; Meuwissen & Choppin, 2015).

As an advocate and de facto cosponsor of a large-scale high-stakes assessment that determines whether someone can become a teacher, AACTE has a responsibility to monitor and follow up on credible criticisms and challenges leveled at the assessment. However, to our knowledge, the organization has so far opted to avoid confronting such concerns or acting on the critiques we have raised. Ironically, reference to the *AERJ* article currently appears in just one document found on the the AACTE website: the eight-page response from SCALE and Pearson. As far as we have been able to determine, AACTE has made no other mention of the article and its claims in any organizational communication at all, including its daily blog, *EdPrepMatters*. Nor has the organization contacted us to discuss the article in any kind of forum.

Response from states

No states reached out to us for any type of discussion with regard to the edTPA. In an effort to understand why, we used the Freedom of Information Act to request communications about edTPA after December 2019 between Pearson/SCALE and states that use the assessment. In state after state, we saw almost no proactive steps taken to address the concerns raised by the article. In fact, in most of the states from which we received information, we did not find a single documented question from state representatives to Pearson or SCALE.

In several states, someone with technical expertise wrote to SCALE or Pearson and said that the claims in the article appeared sound and that they awaited a response. For example, in a communication from Connecticut to SCALE and Pearson, the state official said:

> I just finished reading the Gitomer, et al paper . . . I have to say, from a measurement perspective, it’s a strong paper . . .
>
> Any idea when your rebuttal will be released? Earlier this fall, Connecticut legislators formed a task force to discuss and make recommendations about edTPA in Connecticut, due to EPP [Education Preparation Provider] faculty pushback (e.g., letters and calls). The task force has met once already and will meet again very soon. We will really need a strong paper from SCALE/Pearson in response to the Gitomer article.

In a follow-up message, the official said, “I should also mention that although we do have push back from specific EPP faculty, we also have support for edTPA from others!”

Notably, we received no documented evidence from the states of any further questions about edTPA following the release of *Affirming the Validity and Reliability of edTPA*. For example, the state of Illinois initially posed a number of questions referring to concerns raised in our paper, but we saw no evidence that they reiterated those questions after Pearson and SCALE issued their eight-page response. In general, state personnel appeared to give priority to tamping down our concerns and protecting the edTPA rather than addressing those concerns and protecting aspiring educators from being denied a teaching license unfairly.

For the most part, states do not engage in the same kind of technical review for teacher assessments as they do for student assessments. For example, states have to submit detailed plans addressing the technical quality of their student assessments to the U. S. Department of Education, but no such requirement exists with regard to teacher assessments. Facing policy mandates and lacking resources for a thorough technical review, states often find themselves with little more to go on than the evidence and claims provided by the test developers. However, while we understand the limitations states face, we were troubled that their communications appear to reflect more concern about maintaining the current system than about holding vendors to professional standards. We suspect that they are not eager to change testing programs after having worked hard to embed them in their systems.

States could have approached this in other ways. They could have interrogated our initial claims and, if they determined our claims to be correct, tried to figure out what harm has been done to teaching candidates and teacher education programs. They could have demanded additional evidence or analysis from SCALE to evaluate the appropriateness of specific passing scores, or to determine how much confidence should be placed in the edTPA scores. In short, they could have made an effort (and still can) to examine whether the edTPA was effective in ensuring the quality of the teaching force and whether it was doing so in a way that was fair to prospective teachers.

The implications of looking away

We stand by our concerns about the edTPA and will do so until independent technical experts refute our claims. However, egregious practices like those we have identified can only be addressed if relevant stakeholders demonstrate the will to protect the interests of those affected. Both AACTE and the states can push the developers of edTPA to improve their scoring and reporting so that it meets basic professional standards of measurement practice. Instead, they have so far chosen to give Pearson and SCALE a pass by accepting an unvetted document that satisfies bureaucratic and political needs but does nothing to ensure the integrity of accountability systems in which they provide leadership roles.

The edTPA response to our criticism provides a clear example of what political scientists Peter Bachrach and Morton Baratz (1963) called “nondecision-making.” Bachrach and Baratz considered decision making in the context of power relationships and argued that:

> Many investigators have also mistakenly assumed that power and its correlatives are activated and can be observed only in decision-making situations. They have overlooked the equally, if not more important area of what might be called “nondecision-making”, *i.e.*, the practice of limiting the scope of actual decision-making to “safe” issues by manipulating the dominant community values, myths, and political institutions and procedures. To pass over this is to neglect one whole “face” of power. (p. 632)

Studying how decisions are made only sheds light on one part of the policy process. Equally important are the *non*decisions of policy makers. Nondecision making is not equivalent to making a decision *not to do something*. Rather, it involves a choice *not to engage with or act* on an issue or set of concerns in any significant way. While the reasons for this nondecision may or may not be clearly articulated, the result is the preservation of a certain policy, status quo, or power relation that is implicitly deemed preferable to the alternatives.

As researchers who work in measurement, assessment, and teacher professional development, we do not intend this piece as a general criticism of portfolio assessments for teacher certification. That discussion entails a complex interplay of technical and policy issues and priorities that are certainly worthy of much further analysis. But no matter where one comes down on this issue, there is no place for high-stakes assessments that do not satisfy minimum standards of measurement. With respect to edTPA, the states and AACTE have engaged in nondecision making, privileging the interests of the assessment organizations, not those of teachers or the American public.

References

Bachrach, P. & Baratz, M.S. (1963). Decisions and nondecisions: An analytical framework. *American Political Science Review, 57* (3), 632-642.

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*.

Greenblatt, D. (2016). The consequences of edTPA. *Educational Leadership, 73* (8), 51-54.

McWeeney, D.M. (2020). *Working group to study issues relating to the implementation of the pre-service assessment, edTPA, as adopted by the Connecticut State Board of Education, December 7, 2016: Final report, January 31, 2020.* Hartford, CT: Connecticut General Assembly.
https://doi.org/ 10.13140/RG.2.2.33381.42723

Meuwissen, K.W. & Choppin, J.M. (2015). Preservice teachers’ adaptations to tensions associated with the edTPA during its early implementation in New York and Washington states. *Education Policy Analysis Archives, 23* (103).

Pecheone, R.L., Whittaker, A., Klesch, H., & Esbenshade, L. (2018, April). *Examining the impact of beginning teacher assessment & accountability systems (edTPA): Summary of validity and reliability studies for the edTPA: January 1, 2015-January 1, 2016*. Presented at the Annual Meeting of the National Council of Measurement in Education, New York, NY.

Stanford Center for Assessment, Learning, and Equity (SCALE). (2018). *Educative assessment & meaningful support: 2017 edTPA administrative report.* Stanford, CA: Author.

Note: This is an evolving story. Since this article was edited and approved, Drew Gitomer received an invitation from AACTE’s *Journal of Teacher Education* to participate in a session about the edTPA at the journal’s annual conference.

Drew H. Gitomer Jose Felipe Martinez Dan Battey

DREW H. GITOMER (drew.gitomer@gse.rutgers.edu; @DrewGitomer) is Rose & Nicolas DeMarzo Chair in Education at Rutgers University, New Brunswick, NJ. He is the editor, with Courtney A. Bell, of the *Handbook of Research on Teaching* (5th ed., AERA, 2016).