Journal of Teacher Education
2017, Vol. 68(4) 377–393

DOI: 10.1177/0022487117702582
journals.sagepub.com/home/jte
SAGE

Evaluating Prospective Teachers: Testing the Predictive Validity of the edTPA

Dan Goldhaber[1], James Cowan[1], and Roddy Theobald[1]

Abstract

We use longitudinal data from Washington State to provide estimates of the extent to which performance on the edTPA, a performance-based, subject-specific assessment of teacher candidates, is predictive of the likelihood of employment in the teacher workforce and value-added measures of teacher effectiveness. While edTPA scores are highly predictive of employment in the state's public teaching workforce, evidence on the relationship between edTPA scores and teaching effectiveness is more mixed. Specifically, continuous edTPA scores are a significant predictor of student mathematics achievement in some specifications, but when we consider that the edTPA is a binary screen of teaching effectiveness (i.e., pass/fail), we find that passing the edTPA is significantly predictive of teacher effectiveness in reading but not in mathematics. We also find that Hispanic candidates in Washington were more than 3 times more likely to fail the edTPA after it became consequential in the state than non-Hispanic White candidates.

Keywords

teacher education preparation, certification/licensure, value added

Background: The Teacher Education Accountability Movement

It is fair to say that teacher education programs (TEPs) are facing significant scrutiny over the inservice performance of their graduates. About 75% of the roughly 100,000 novice teachers who enter the public school workforce each year are trained in a traditional college or university setting, and there is significant policy concern that the preparation that prospective teachers receive is not adequate to ensure they are ready to teach on their first day in a classroom. Former Education Secretary Arne Duncan, for instance, stated, "By almost any standard, many if not most of the nation's 1,450 schools, colleges and departments of education are doing a mediocre job of preparing teachers for the realities of the 21st century classroom" (U.S. Department of Education, 2009, para. 3).

Given this environment, it is not surprising that there are a number of new initiatives designed to hold TEPs more accountable, either through direct measures of the training they provide teacher candidates or based on output measures, such as the value added of candidates who enter the teaching workforce. One of the ways that TEPs and states have responded to this increased accountability pressure is by adopting the edTPA, a performance-based, subject-specific assessment that is administered to teacher candidates during their student teaching assignment. There has been remarkably rapid policy diffusion of this assessment from its initial field testing in 2012 to full implementation (Gottlieb, Hutt, & Cohen, 2016): The edTPA is now used by more than 600 TEPs in 40 states, and passing the edTPA is a requirement for licensure in seven states.[1] Yet, despite the rapid adoption of this assessment, critics of the edTPA (e.g., Greenblatt & O'Hara, 2015) point out that there is limited large-scale research linking edTPA scores to outcomes for inservice teachers and their students.

There are several theories of action for how teacher performance assessments like the edTPA might improve the quality of the teacher workforce. First, the edTPA can be used as a high-stakes screen and "provide a consistent standard for entry into the profession" (Hill, Hansen, & Stumbo, 2011); this is how the edTPA is currently used in states in which the assessment is a requirement to participate in the labor market.[2] This use of the edTPA requires predictive validity around the cut point adopted for labor market participation, which is set to different scores in different states through "standard setting conferences" described in edTPA (2015).[3]

The edTPA might also improve the quality of the teaching workforce by affecting candidate teaching practices. Indeed, the edTPA is described by its developers as an "educative assessment" that "supports candidate learning and preparation program renewal" (edTPA, 2015), and Hill et al.

[1]CALDER, American Institutes for Research, Seattle, WA, USA

Corresponding Author:
Roddy Theobald, CALDER, American Institutes for Research, 3876 Bridge Way N., Suite 201, Seattle, WA 98103, USA.
Email: rtheobald@air.org

(2011) suggest that the teacher performance assessments like the edTPA could "describe expectations for novice teaching and set a trajectory of improvement over the developmental continuum" (p. 6). This could occur at the individual teacher candidate level if, for instance, participation in the edTPA directly influences the teaching practices of teacher candidates. Alternatively, this could occur at the TEP level if, for instance, participation in the edTPA influences the training provided by TEPs. Finally, the edTPA might be used for hiring purposes; for instance, school systems might be more likely to hire teacher applicants with higher edTPA scores. Each of these potential mechanisms for workforce improvement requires that the edTPA provides a signal of quality teaching, that is, that there is predictive validity away from the cut point such that differences in edTPA performance (at the candidate or institution level) might be indicative of teacher quality.

In this article, we use longitudinal data from Washington State that includes information on teacher candidates' scores on the edTPA to provide estimates of the extent to which edTPA scores are predictive of the likelihood of entry into the teacher workforce and value-added measures of teacher effectiveness (i.e., predictive validity). Specifically, we test different theories of action for how the edTPA might improve the quality of the teacher workforce by considering the predictive validity of the edTPA as both a screen and a signal of future teacher effectiveness.

Despite the fact that the edTPA was not consequential for some of the teacher candidates in our sample, we find that edTPA scores—both in terms of passing status and continuous scores—are highly predictive of the probability that a teacher candidate is employed the following year in the state's public teaching workforce. Evidence on the connection between performance and value-added measures of teacher effectiveness is more mixed. When we consider the edTPA as a binary screen of teaching effectiveness (i.e., pass/fail), we find that passing the edTPA is significantly predictive of teacher effectiveness in reading but not in mathematics. Continuous edTPA scores provide a signal of future teaching effectiveness in mathematics in some specifications, but are not statistically significant in reading. In both reading and mathematics, the relationship between continuous edTPA scores and teacher effectiveness is somewhat stronger for candidates who took the test after it became consequential in Washington, suggesting that the edTPA may provide a better signal of teacher quality when stakes are attached to the scores.

We also find that Hispanic teacher candidates score far lower than non-Hispanic White candidates on the assessment. In fact, Hispanic candidates in Washington were more than 3 times more likely to fail the edTPA after it became consequential in the state than non-Hispanic White candidates (13.7% for Hispanic candidates compared with 3.7% for non-Hispanic White candidates). This difference in passing rates strongly implies that the high-stakes use of the edTPA in Washington may have an adverse impact on the diversity of the state's teacher candidate pool. However, it is important to be cautious about interpreting this as an effect on the diversity of the state's teacher workforce. It is possible, for example, that teachers who fail the test would be unlikely to obtain teaching positions in the absence of the edTPA requirement or that the high-stakes use of the edTPA elicits other behavioral changes that affect who pursues a career as a teacher.

The rest of the article proceeds as follows: In "Assessment of Prospective Teachers and the Role of the edTPA" section, we provide additional information regarding teacher licensure and the edTPA in particular. We describe our data and analytic approach in "Data and Analytic Approach" section, present our findings in "Results" section, outline some extensions in "Policy Implications" section, and offer concluding remarks in "Conclusion" section.

Assessment of Prospective Teachers and the Role of the edTPA

There are various ways that teacher candidates are typically assessed and judged to be eligible—that is, licensed—to teach in public schools. Licensure in many states requires that prospective teachers graduate from an approved TEP and complete some preservice student teaching, although the last decade has also seen an increased reliance on teachers entering the profession through state-approved alternative routes. Forty-nine of 50 states also require potential teachers to pass licensure tests that cover basic skills, content knowledge, and/or professional knowledge.

The edTPA, by design, is quite different from traditional question-and-answer licensure tests: It is a portfolio-based, subject-specific assessment akin to the National Board for Professional Teacher Standards (NBPTS) assessment of inservice teachers. The edTPA was initially developed by researchers at Stanford University's Center for Assessment, Learning, and Equity (SCALE) and has been further developed and distributed through a partnership between SCALE, the American Association of Colleges for Teacher Education (AACTE), and Evaluation Systems (a member organization of the Pearson Education group). The edTPA was initially introduced in two large-scale field tests in 2011-2012 and 2012-2013 and was "operationally launched" in 2013-2014 (Pecheone, Shear, Whittaker, & Darling-Hammond, 2013). The edTPA relies on the scoring of teacher candidates who are videotaped while teaching three to five lessons from an instructional unit to one class of students, along with assessments of teacher lesson plans, student work samples and evidence of student learning, and reflective commentaries by the candidate. Candidates pay a US$300 fee to take the edTPA and often take several months to prepare their portfolios for submission (e.g., Jette, 2014).

The edTPA is a subject-specific assessment with different versions aligned with 27 different teaching fields (e.g., "Early

Childhood," "Secondary Mathematics," etc.).[4] Each of these versions of the edTPA contains 15 different rubrics, each of which is scored on a 1 to 5 scale; the rubrics have equal weight so the range of possible summative scores (for tests with no incomplete rubric scores) is 15 to 75.[5] The 15 rubrics that are used to calculate a candidate's summative score in Washington State are grouped into three areas: Planning (e.g., "Planning for Subject-Specific Understandings"), Instruction (e.g., "Engaging Students in Learning"), and Assessment (e.g., "Analysis of Student Learning").[6] Teacher candidates in Washington State are also scored on three additional student voice rubrics (e.g., "Eliciting Student Understanding of Learning Targets"), which are designed to incorporate student-produced material into a teacher's evaluation. For reasons discussed in the next section, these rubric scores are not currently used in computing a candidate's summative score.[7]

Proponents of the edTPA argue that the assessment and its precursors are authentic measurement tools that can be used to predict teacher candidates' success in the classroom (e.g., Darling-Hammond, 2009; edTPA, 2015; Hill et al., 2011). While the edTPA is designed to assess individual teacher candidates, it is also thought to inform improvements in TEPs. Some states are, in fact, using the average edTPA performance of teacher candidates at an institution as a measure of institutional quality and/or in the accreditation process. In addition, the use of the edTPA is heavily promoted by AACTE, which touts the assessment as a means of improving " . . . the information base guiding the improvement of teacher preparation programs [and] strengthen[ing] the information base for accreditation and evaluation of program effectiveness."[8]

Claims about the potential predictive validity of the edTPA are based on a small literature demonstrating that *inservice* teacher performance on portfolio-based assessments like the NBPTS assessment (Cantrell, Fullerton, Kane, & Staiger, 2008; Cowan & Goldhaber, 2016; Goldhaber & Anthony, 2007) and Washington State's ProTeach assessment (Cowan & Goldhaber, 2014) are predictive of teacher effectiveness, as well as two small-scale pilot studies of the edTPA's precursor, the Performance Assessment for California Teachers (PACT).[9] Specifically, Newton (2010) finds positive correlations between PACT scores and future value-added for a group of 14 teacher candidates, while Darling-Hammond, Newton, and Chung Wei (2013) use a sample of 52 mathematics teachers and 53 reading teachers and find that a one standard deviation increase in PACT scores is associated with a .03 standard deviation increase in student achievement in either subject.[10] Beyond the fact that these estimates are based on small sample sizes, however, there are several substantive differences between the edTPA and PACT in terms of scoring, implementation, and standards alignment.[11]

As described in the next section, the administrative data we utilize for our research allows us to leverage a larger sample size of teachers (over 200 in both mathematics and reading) than the PACT studies cited above. Each of these teachers took the edTPA after its full national implementation in the 2013-2014 school year. It is important to note, however, that the edTPA did not become consequential in Washington State until January 2014,[12] so candidates who failed the test in fall 2013 (as well as candidates who failed after January 2014 but subsequently retook and passed the test) provide an opportunity to observe candidates who failed the test but still entered the public teaching workforce.

While this study is one of the first to provide evidence on the validity of the edTPA as a measure of classroom performance, it is important to distinguish the validity of the edTPA as an assessment of teaching practice from its efficacy as a teacher licensing tool. In particular, while validity is a significant prerequisite for using the edTPA to support effective licensure policy, extrapolating from these results to the effects of particular policies requires imposing additional assumptions beyond those that we test here.

In particular, four features of common licensure policies limit such additional conclusions. First, licensure policies may change the population of potential teachers if candidates view the test as costly. There is some evidence from changes to state licensing provisions that licensure tests discourage some candidates with high academic achievement or outside wage offers from pursuing teaching as a profession, although evidence on overall effects on student achievement is mixed (Angrist & Guryan, 2008; Larsen, 2015; Wiswall, 2007). Second, policies typically allow candidates to attempt the assessment multiple times. In the second half of the 2013-2014 school year (when the edTPA was consequential), 4% of test takers failed the edTPA the first time they took it, but about half of these candidates eventually passed the test. Third, the matching of teacher candidates to teaching positions may provide additional screening beyond what is required by law. For example, it is not clear that the small number of teachers in our sample who never pass the edTPA would obtain employment even in the absence of testing requirements. Finally, licensure systems like the edTPA might have systemwide effects on teacher quality. If participation in the edTPA raises overall performance, the signaling effects we estimate here may understate the overall effects of implementing testing requirements. The policy effects of national implementation of the edTPA, and similar authentic licensure assessments, therefore remains an important area for future research.

Data and Analytic Approach

Data

Our research uses administrative data on teacher candidates provided by Washington State's Professional Educator Standards Board (PESB), as well as data on Washington State students, teachers, and schools maintained by the Office of the Superintendent of Public Instruction (OSPI).

The PESB data includes scores on each individual edTPA rubric (as well as the final summative score) for *all* teacher candidates who took the edTPA in Washington State, not just those who ultimately are employed in the teacher workforce. As described in the previous section, the 15 rubrics used to compute the summative score can be combined into three subscores: Planning (Rubrics 1-5), Instruction (Rubrics 6-10), and Assessment (Rubrics 11-15).[13]

Washington State participated in the edTPA field test in the 2012-2013 school year (see Pecheone et al., 2013) and the PESB data include teacher candidate scores from this pilot year and two subsequent school years (2013-2014 and 2014-2015) after the full national implementation of edTPA. Because there were substantive changes to the assessment between the pilot year and full implementation (edTPA, 2015) and because inservice data are not yet available for teacher candidates who took the edTPA in 2014-2015, our primary results focus on the 2,362 teacher candidates from Washington State TEPs who took the edTPA in the 2013-2014 school year. In most cases, we consider edTPA scores from each candidate's first test administration; although in cases where a candidate received an incomplete score and subsequently resubmitted his or her materials within a month, we disregard the initial incomplete score and consider a candidate's subsequent submission.[14]

We link these edTPA scores to data from OSPI that include test scores on other licensure tests that teacher candidates must also pass to be eligible to teach, such as the Washington Educator Skills Test–Basic (WEST-B), an assessment of basic skills in reading, writing, and mathematics that has been a requirement for admission into Washington State TEPs since 2002.[15] Among teacher candidates in the edTPA sample, 60.29% entered the state's public teaching workforce in the 2014-2015 school year (defined as being employed in a certificated teaching position), and for these 1,424 teacher candidates, the OSPI data also include information about their school assignments, race, gender, and ethnicity.

For the subset of 277 teacher candidates who enter the workforce and teach mathematics or reading in Grades 4 to 8 (i.e., grades and subjects in which both current and prior test scores are available or the value-added sample), we can investigate the relationship between edTPA performance and student achievement. Specifically, we observe annual student test scores in mathematics and reading in Grades 3 to 8 (also provided by OSPI) on the state's Measures of Student Progress (MSP) examination in 2012-2013 and 2013-2014 and Smarter Balanced Assessment (SBA) in the 2014-2015 school year.[16] We standardize these scores within grade and year and connect them to additional student demographic information (gender, race/ethnicity, special education status, free/reduced-priced lunch eligibility, and English learner status), and through a unique link in the state's Comprehensive Education Data and Research System (CEDARS) data system, to data on the student's teachers in mathematics and reading (described above).[17]

Table 1 summarizes data for prospective teachers who took the edTPA assessment in 2013-2014 for all candidates (columns 1-6) and for candidates who appear in the teaching workforce in 2014-2015 (columns 7-12). Within each set of columns, we present summary statistics for all individuals within the group (columns 1 and 7) and by quintile of performance on the edTPA (columns 2-6 and 8-12).[18] In column 1, we see that the overall first time pass rate on the test, 93.9%, was quite high because Washington State had set a low cut score of 35, but this passing rate would have been only 86.5% had the state used its future cut score of 40.

The summary statistics for teacher candidates by quintile of performance on the edTPA (columns 2-6) make it clear that there is a correlation between edTPA performance and the WEST-B basic-skills licensure tests that are required for entry into Washington State's TEPs.[19] It is also immediately clear that teachers who perform better on the edTPA are more likely to be employed in Washington State's public schools in the subsequent year: Only 50.8% of first quintile (lowest quintile) teachers are observed teaching versus 64.6% of fifth quintile (top quintile) teachers. We also observe large differences in performance between Hispanic and non-Hispanic White teacher candidates. Specifically, Hispanic candidates are about twice as likely to score in the lowest quintile of the edTPA as in the middle three quintiles and 4 times as likely to score in the lowest quintile as in the top quintile.[20]

We further explore the differences in edTPA performance by teacher candidate race/ethnicity in Table 2 and Figure 1. Hispanic teacher candidates score significantly lower than non-Hispanic White candidates on the total score, all three subscores, and all 15 individual rubrics.[21] In addition, Hispanic candidates in Washington were more than 3 times more likely to fail the edTPA after it became consequential in the state than non-Hispanic White candidates: 13.7% of Hispanic candidates failed the test after January 2014, compared with 3.7% of non-Hispanic White candidates.[22] Although this difference in passing rates suggests that the high stakes use of the edTPA in Washington may adversely affect the state's teacher workforce diversity, we do not find that that first-year teachers in the 2014-2015 school year (the year after the edTPA became consequential) are less diverse than in earlier years; in fact, 7.39% of all first-year teachers in 2014-2015 are Hispanic, compared with 4.47% in 2013-2014. It is also unclear whether the high-stakes use of the edTPA elicits other behavioral changes that affect who pursues a career as a teacher or whether Hispanic teachers may be more likely to receive emergency credentials to teach in high-needs areas like in English Language Learner programs.

Analytic Approach

To investigate the relationship between edTPA scores and the probability of workforce entry, we first define p_{jkt} as the probability that teacher candidate j who took edTPA test type k in 2013-2014 appears as a Washington State public schoolteacher

Table 1. Summary Statistics.

	Teacher candidate sample						Teacher sample					
	1	2	3	4	5	6	7	8	9	10	11	12
Variable	All	Q1	Q2	Q3	Q4	Q5	All	Q1	Q2	Q3	Q4	Q5
Total score	46.339 (6.960)	37.272 (4.535)	44.100 (0.804)	47.045 (0.813)	50.398 (1.133)	56.204 (3.069)	46.960 (6.726)	37.537 (4.398)	44.122 (0.797)	47.035 (0.814)	50.368 (1.124)	56.240 (3.003)
Planning subscore	15.924 (2.644)	12.962 (2.216)	15.218 (1.234)	16.138 (1.256)	17.284 (1.410)	19.102 (1.624)	16.102 (2.584)	13.088 (2.182)	15.135 (1.297)	16.095 (1.304)	17.243 (1.458)	19.113 (1.628)
Instruction subscore	15.460 (2.416)	12.869 (1.771)	14.783 (1.203)	15.515 (1.121)	16.473 (1.389)	18.623 (1.678)	15.661 (2.334)	12.974 (1.733)	14.857 (1.133)	15.574 (1.132)	16.465 (1.395)	18.579 (1.642)
Assessment subscore	14.922 (2.948)	11.395 (2.317)	14.055 (1.293)	15.361 (1.286)	16.605 (1.291)	18.473 (1.638)	15.165 (2.896)	11.432 (2.317)	14.094 (1.317)	15.327 (1.350)	16.627 (1.287)	18.542 (1.603)
% passing WA score (35)	0.939	0.743	1.000	1.000	1.000	1.000	0.951	0.759	1.000	1.000	1.000	1.000
% passing future score (40)	0.865	0.436	1.000	1.000	1.000	1.000	0.889	0.456	1.000	1.000	1.000	1.000
WEST-B Reading	271.016	266.396	271.200	271.656	272.459	274.659	271.482	265.681	271.269	271.901	273.364	275.667
	(16.139)	(17.522)	(16.709)	(15.307)	(14.961)	(14.221)	(15.997)	(17.806)	(17.118)	(14.711)	(14.623)	(13.276)
WEST-B Writing	264.340	257.273	264.049	265.408	267.332	269.685	264.910	256.710	264.166	266.214	267.905	270.266
	(18.049)	(19.124)	(17.832)	(17.259)	(15.523)	(17.224)	(17.550)	(18.740)	(17.113)	(16.468)	(15.455)	(16.578)
WEST-B Math	279.548	274.924	280.000	280.233	281.775	282.064	280.093	274.978	280.261	280.838	282.431	282.357
	(17.649)	(20.092)	(16.231)	(16.797)	(15.768)	(17.452)	(17.288)	(18.889)	(16.128)	(15.973)	(15.946)	(18.206)
Female	0.764	0.745	0.732	0.785	0.792	0.779	0.768	0.762	0.721	0.791	0.773	0.797
White	0.785	0.756	0.778	0.796	0.785	0.817	0.779	0.759	0.755	0.789	0.791	0.804
Asian	0.045	0.028	0.056	0.040	0.058	0.049	0.050	0.033	0.063	0.042	0.060	0.054
Black	0.012	0.011	0.014	0.011	0.014	0.012	0.015	0.010	0.022	0.011	0.020	0.014
Hispanic	0.060	0.114	0.040	0.047	0.058	0.028	0.063	0.127	0.050	0.053	0.050	0.034
Multirace	0.046	0.039	0.044	0.063	0.044	0.044	0.044	0.026	0.041	0.067	0.040	0.047
Entering workforce	0.602	0.508	0.599	0.615	0.674	0.646	1.000	1.000	1.000	1.000	1.000	1.000
Reading VAM sample	0.088	0.058	0.086	0.121	0.095	0.091	0.139	0.107	0.135	0.190	0.136	0.132
Math VAM sample	0.087	0.072	0.084	0.110	0.095	0.077	0.137	0.134	0.132	0.173	0.136	0.111
Teacher candidates	2,376	569	501	447	432	427	1,508	307	319	284	302	296

Note. We omit summary statistics for American Indian, Alaskan Native, and Other/unspecified race candidates due to small cell sizes. Four hundred thirteen teacher candidates and 185 teachers are missing WEST-B scores, with the distribution of missing scores relatively uniform across quintiles. Standard deviations of continuous variables in parentheses. VAM = value-added model; WEST-B = Washington Educator Skills Test–Basic.

Table 2. edTPA Performance by Teacher Candidate Race and Ethnicity.

edTPA performance	Teacher candidate sample					
	Overall	Asian	Black	White	Hispanic	Multirace
Total score	46.339 (6.960)	47.602† (6.379)	46.966 (5.698)	46.544 (6.901)	42.972*** (7.747)	46.582 (6.454)
Planning subscore	15.924 (2.644)	16.389† (2.537)	16.034 (1.927)	15.966 (2.618)	14.913*** (2.930)	16.250 (2.530)
Instruction subscore	15.460 (2.416)	15.593 (2.173)	15.983 (2.230)	15.515 (2.427)	14.559*** (2.448)	15.445 (2.305)
Assessment subscore	14.922 (2.948)	15.583* (2.666)	14.845 (2.435)	15.031 (2.910)	13.451*** (3.452)	14.868 (2.698)
Overall % passing WA score (35)	0.939	0.972†	1.000***	0.943	0.839**	0.964
Passing score (35): Preconsequential	0.876	0.812	1.000***	0.878	0.780	1.000***
Passing score (35): Postconsequential	0.958	1.000***	1.000***	0.964	0.863**	0.956
Overall % passing future score (40)	0.865	0.926†	0.897	0.874	0.706***	0.864
Passing score (40): Preconsequential	0.769	0.625	1.000***	0.792	0.634†	0.700
Passing score (40): Postconsequential	0.895	0.978***	0.889	0.901	0.735***	0.900
Planning: Planning for subject-specific understanding	3.250 (0.658)	3.361 (0.673)	3.190 (0.451)	3.263 (0.653)	3.010*** (0.670)	3.341 (0.628)
Planning: Planning to support varied learning needs	3.199 (0.724)	3.278 (0.635)	3.328 (0.602)	3.201 (0.726)	3.052* (0.738)	3.305 (0.739)
Analyzing teaching: Using knowledge of students to inform teaching and learning	3.215 (0.691)	3.306 (0.733)	3.172 (0.602)	3.217 (0.685)	3.094† (0.717)	3.277 (0.676)
Academic language: Identifying and supporting language demands	3.114 (0.651)	3.162 (0.641)	3.017 (0.491)	3.136 (0.642)	2.857*** (0.740)	3.077 (0.618)
Planning: Planning assessments to monitor and support student learning	3.146 (0.711)	3.282* (0.639)	3.328 (0.631)	3.150 (0.709)	2.899*** (0.817)	3.250 (0.652)
Instruction: Learning environment	3.251 (0.524)	3.231 (0.570)	3.414 (0.552)	3.258 (0.517)	3.157* (0.446)	3.245 (0.589)
Instruction: Engaging students in learning	3.104 (0.618)	3.139 (0.579)	3.224 (0.560)	3.123 (0.620)	2.916*** (0.602)	3.091 (0.606)
Instruction: Deepening student learning	3.060 (0.661)	3.065 (0.552)	3.086 (0.584)	3.068 (0.666)	2.874** (0.723)	3.091 (0.606)
Instruction: Subject-specific pedagogy: Using representations	3.076 (0.668)	3.167 (0.580)	3.138 (0.533)	3.084 (0.677)	2.892** (0.680)	3.064 (0.639)
Analyzing teaching: Analyzing teaching effectiveness	2.968 (0.724)	2.991 (0.652)	3.121 (0.764)	2.983 (0.726)	2.720*** (0.740)	2.955 (0.634)
Assessment: Analysis of student learning	3.148 (0.761)	3.370** (0.718)	3.138 (0.625)	3.168 (0.745)	2.790*** (0.901)	3.123 (0.668)
Assessment: Providing feedback to guide learning	3.165 (0.782)	3.236 (0.715)	3.155 (0.780)	3.192 (0.784)	2.860*** (0.829)	3.182 (0.735)
Assessment: Student use of feedback	2.711 (0.760)	2.801 (0.752)	2.707 (0.575)	2.738 (0.760)	2.472*** (0.843)	2.550** (0.717)
Academic Language: Analyzing student's language use and subject-specific learning	2.830 (0.696)	2.880 (0.615)	2.724 (0.544)	2.852 (0.691)	2.538*** (0.822)	2.909 (0.668)
Analyzing teaching: Using assessment to inform instruction	3.068 (0.785)	3.296** (0.680)	3.121 (0.715)	3.080 (0.774)	2.790*** (0.871)	3.105 (0.824)
Observations	2,376	108	29	1,864	143	110

Note. We omit summary statistics for American Indian, Alaskan Native, and Other/Unspecified race candidates due to small cell sizes. Significance stars are from a two sample *t* test with unequal variances between white teacher candidates and the race indicated by column. Passing rates for both the preconsequential period and for the future cut score assume no other behavioral changes are associated with the change of cut score or stakes attached to the test.

$^{\dagger}p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

in the 2014-2015 school year and estimate a simple logit model for all 2,238 teacher candidates in the sample:

$$\log\left(\frac{p_{jk}}{1-p_{jk}}\right)=\alpha_0+\alpha_1\text{TPA}_{jk}+\alpha_k+\varepsilon_{jk}. \quad (1)$$

In the base specification of the model in Equation 1, TPA_{jk} is a binary variable indicating whether teacher candidate j passed the edTPA on the first test sitting. Given that all specifications include fixed effects for test type k, all coefficients can be interpreted as relative to other teacher candidates who took the same test type.[23] Although the coefficient of interest α_1 is on the log odds scale, we present all estimates as average marginal effects. We also estimate three other specifications of the model in Equation 1 in which (a) TPA_{jk} is an indicator for whether candidate j *would have* passed the edTPA at the state's future (and higher) cut score; (b) TPA_{jk} is a continuous variable indicating the edTPA score of candidate j (standardized relative to all test takers); and (c) $\mathbf{TPA}_{jk}$ is a vector of scores for candidate j across the three subscores on the test (each standardized relative to all test takers).

To investigate the predictive validity of the edTPA in terms of predicting the achievement of a teacher candidate's future students, we estimate value-added models (VAMs) intended to separate the impact of teacher characteristics (such as edTPA scores) from other variables that influence student test performance (see Koedel, Mihaly, & Rockoff, 2015 for review). Specifically, we estimate variants of the following VAM only for the candidates who enter the teaching workforce and are linked to current and lagged student achievement data (204 in reading, 206 in mathematics):

$$\begin{aligned}Y_{ijgkst}=&\beta_0+\beta_1Y_{i,t-1}+\beta_2X_{it}+\beta_3C_{ist}+\beta_4Z_{jt}\\&+\beta_5\text{TPA}_{jk}+\beta_g+\beta_k+\varepsilon_{ijkgst}.\end{aligned} \quad (2)$$

In Equation 2, Y_{ijgkst} is the SBA score of student i in grade g, subject s, and year t (the 2014-2015 school year for all students), while in the classroom of teacher j who took edTPA test type k. $\mathbf{Y}_{i,t-1}$ is a vector of student i's prior year test scores in mathematics and reading. The student test scores in both Y_{ijgst} and $Y_{i,t-1}$ are standardized by test, grade, and year across all test takers. Therefore, the units of the coefficients on the right side of Equation 2 are standard deviations of student performance (relative to other scores on the same test in the same grade and year). $\mathbf{X}_{it}$ is a vector of student covariates for student i, in year t, which includes indicators for race/ethnicity, gender, free or reduced-priced lunch eligibility, gifted/highly capable, limited English proficiency (LEP), special education, and learning disabled. $\mathbf{C}_{ist}$ is a vector of aggregated student characteristics in the student's classroom, while Z_{jt} an indicator for whether or not a teacher possesses an advanced degree in year t.[24] All specifications include fixed effects for grade g and test type k, so all results can be interpreted as relative to other students in the same grade whose teachers took the same edTPA test type.

The different specifications of the model in Equation 2 correspond to the different theories of action discussed in the introduction. When we investigate the edTPA as a screening mechanism intended to prevent low-performing teachers from entering the workforce, TPA_{jk} is an indicator for whether candidate j passed the edTPA on the first test administration (or, in a related specification, *would have* passed the edTPA at the state's future cut score). When we investigate the signal value of edTPA scores (i.e., the extent to which a candidate's score could be used as a proxy for future teaching effectiveness), TPA_{jk} is the standardized edTPA score of candidate j (or, in a separate specification, a vector of standardized scores for candidate j across the test's three subscores).

We estimate specifications with only test type fixed effects (the most parsimonious model in which teachers are compared with other teachers who took the same test type), with test type and TEP fixed effects (in which teachers are compared with other teachers who took the same test type and graduated from the same TEP), and with test type and school district fixed effects (in which teachers are compared with other teachers who took the same test type and are teaching in the same school district).[25] We estimate Equation 2 by ordinary least squares (OLS) and cluster standard errors at the teacher level to account for correlation between the errors of students taught by the same teacher.

One challenge in estimating all these specifications is that approximately one third of students in Grades 4 to 8 have missing prior-year test scores because their school participated in Washington State's SBA pilot in the 2013-2014 school year (and the state did not collect their scores). We therefore estimate three types of models: (a) a listwise deletion model that drops all students with missing prior-year test scores (possible in Grades 4-8), (b) an imputation model that uses twice-lagged test scores to impute lagged test scores for students with missing test scores (possible in Grades 5-8), and (3) a stacked model that considers any student with either once-lagged scores, twice-lagged scores, or both and uses missing-value dummies to account for missing data (possible in Grades 4-8). We present primary results from the stacked models because they are based on the largest sample sizes, but estimates from the other models show that the results are not sensitive to these sample considerations.[26]

The broader VAM literature (e.g., Chetty, Friedman, & Rockoff, 2014; Kane, McCaffrey, Miller, & Staiger, 2013) suggests that the VAMs described above account for the potential nonrandom sorting of students to teachers in the sample. A second concern, however, is the potential for sample selection bias. As is the case with other licensure tests, sample selection is a concern if teacher characteristics not captured by the edTPA are relevant for hiring decisions and contribute to teacher effectiveness. The literature on

Table 3. Models Predicting Public Teaching Employment.

	edTPA as a screen				edTPA as a signal			
Variables of interest	1	2	3	4	5	6	7	8
Passing in Washington	0.152***	0.112**						
	(0.042)	(0.042)						
Future Washington passing score			0.137***	0.112***				
			(0.030)	(0.030)				
Total score					0.059***	0.045***		
					(0.011)	(0.011)		
Assessment factor							0.034*	0.030*
							(0.015)	(0.014)
Planning factor							−0.003	−0.004
							(0.014)	(0.014)
Instruction factor							0.035*	0.023
							(0.014)	(0.013)
TEP effects		X		X		X		X
Teachers	2,238	2,238	2,238	2,238	2,238	2,238	2,238	2,238

Note. All models controls for teacher degree level and test type effects. Average marginal effects calculated from logit model in Equation 1. Of the full sample of 2,238 teachers, 2,238 teachers take the same tests with at least one other teacher. Similarly, 2,237 teachers were enrolled in TEPs with at least one other teacher. TEP = teacher education program.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

teacher hiring suggests that this is likely to be the case. For example, administrative and survey evidence suggests that references, interviews, and personality traits are important predictors of employment outcomes, and that several of these measures are related to student achievement (Goldhaber, Grout, & Huntington-Klein, 2014; Harris & Sass, 2014; Jacob, Rockoff, Taylor, Lindy, & Rosen, 2016; Rockoff, Jacob, Kane, & Staiger, 2011). Consequently, teachers who perform poorly in the domains measured by the edTPA but who appear in our sample are likely hired because they possessed some compensating skill or skills that make them more effective teachers. In other words, the candidates we observe with low scores are probably disproportionately high-performing teachers.

We explore this issue empirically in "Results" section below, but we argue that two factors are likely to limit the selection bias in our application. First, we examine the edTPA at a time when it was not fully binding in Washington State. Given the lower cut score and the ability of failing teacher candidates to retake the assessment, the selection probabilities between initial passing candidates and initial failing candidates are not as substantial as they would be if the testing requirement was fully binding.

Second, while nontested teacher skills appear related both to hiring decisions and to teacher effectiveness, this relationship is not particularly strong. For example, analyses of the kinds of subjective data available to hiring authorities suggest that, when combined with observable and objective measures of teacher skill, these measures explain only 10% to 20% of the variation in teacher effectiveness (Goldhaber, Grout, & Huntington-Klein, 2014; Jacob et al., 2016; Rockoff et al., 2011). Results from Jacob et al. (2016) suggest a similar relationship to the probability that a candidate for a position is hired.

Results

In this section, we describe our primary research findings on the extent to which edTPA scores predict: the likelihood of being in the Washington State public teacher workforce (Table 3 and Figure 2), teacher effectiveness in reading (Table 4 and Figure 3), and teacher effectiveness in mathematics (Table 5 and Figure 4). Before discussing our primary findings, however, a few peripheral findings are worth brief mention.[27] In terms of predicting employment in the Washington State teacher labor market, we find both that individual TEPs are associated with different probabilities of employment and that candidates who took the edTPA in a STEM (science, technology, engineering, and mathematics) area are more likely to be employed than are candidates who took an elementary edTPA assessment. Both findings echo earlier results from Goldhaber, Krieg, and Theobald (2014).

When estimating student achievement models, we find that underrepresented minority students (Black and Hispanic), participants in the free and reduced-priced lunch program, and students with reported learning disabilities score lower than their reference groups, all else equal. The magnitudes of these findings are quite similar to what has previously been found in Washington State (e.g., Goldhaber, Liddle, & Theobald, 2013) and other states (e.g., Rivkin, Hanushek, & Kain, 2005). Similar to the employment models, TEPs explain a significant portion of student achievement gains in both mathematics and reading. This finding is similar to evidence from Washington State and other states in terms of the variation in teacher

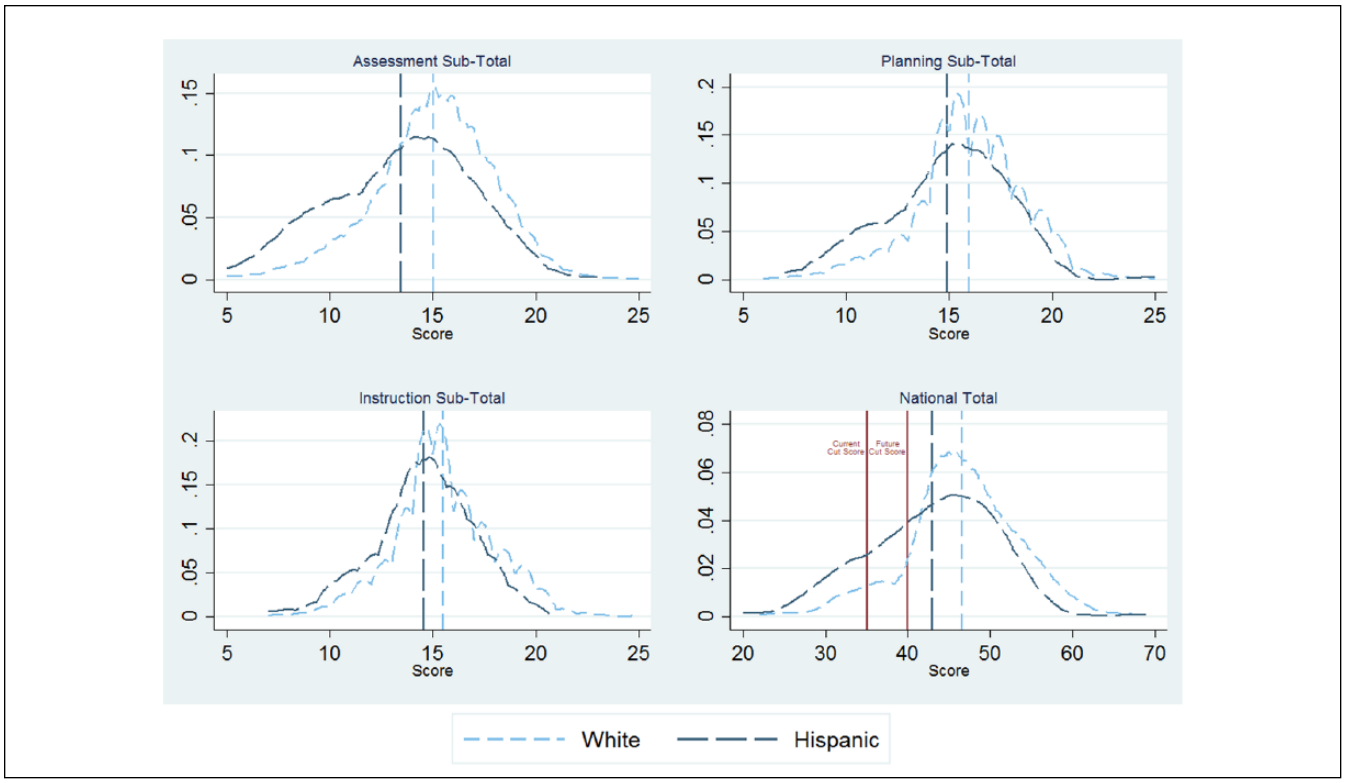

Figure 1. Distribution of edTPA scores for White and Hispanic teacher candidates.

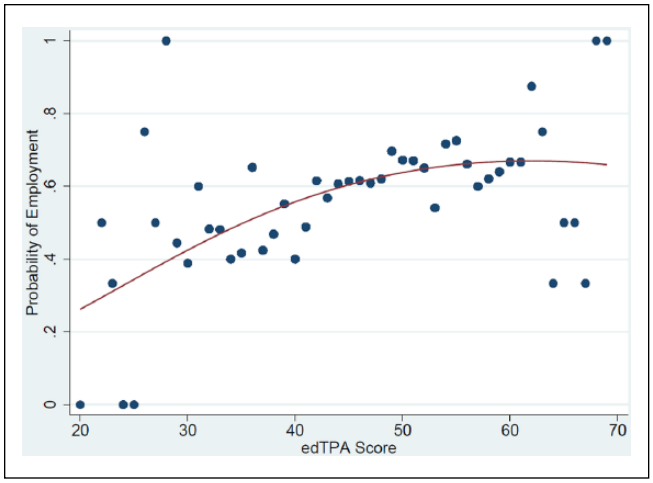

Figure 2. Relationship between edTPA scores and probability of public teaching employment.

effectiveness that can be attributed to TEPs (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Goldhaber et al., 2013; Mihaly, McCaffrey, Sass, & Lockwood, 2013).

edTPA as Predictor of Workforce Entry

Table 3 reports several specifications of models predicting the likelihood of being employed in the Washington State public schoolteacher labor market the year after a candidate takes the edTPA assessment (see Equation 1 above). All coefficients are reported as average marginal effects; so the estimate in column 1, for example, means that teacher candidates who passed the edTPA at the Washington State cut score are 15.2 percentage points more likely to enter the public teaching workforce than are teacher candidates who failed the edTPA at the Washington State cut score, all else equal (i.e., compared with other candidates who took the same test type). The estimated marginal effect is somewhat smaller when candidates are compared with other candidates from the same TEP (column 2) and when we consider candidates who would have passed the test at the future Washington State cut score (columns 3 and 4). These relationships are not surprising given that passing the edTPA is a licensure requirement for some candidates in our sample. Not surprisingly, these relationships are even stronger when we restrict the sample only to teacher candidates who took the edTPA after it became consequential.[28]

Columns 5 to 8 consider continuous measures of edTPA performance as predictors of workforce entry. These continuous scores are standardized across all test takers, so the average marginal effect in column 5 means that a one standard deviation increase in a candidate's edTPA score is associated with a 5.9 percentage point increase in the probability that an average teacher candidate is employed in the teacher

Table 4. Value-Added Results in Reading (Stacked Model).

	edTPA as a screen						edTPA as a signal					
Variables of interest	1	2	3	4	5	6	7	8	9	10	11	12
Passing in Washington	0.251** (0.073)	0.191* (0.080)	0.247*** (0.065)									
Future Washington passing score				0.203** (0.058)	0.149** (0.054)	0.169** (0.058)						
Total score							0.022 (0.017)	0.003 (0.018)	0.006 (0.018)			
Assessment factor										0.031† (0.017)	0.017 (0.016)	0.050* (0.020)
Planning factor										0.020 (0.014)	0.018 (0.014)	−0.007 (0.014)
Instruction factor										−0.025† (0.014)	−0.028† (0.016)	−0.031 (0.019)
TEP effects		X			X			X			X	
District effects			X			X			X			X
Teachers	210	210	210	210	210	210	210	210	210	210	210	210

Note. All models control for student prior performance (either both or just lagged or twice lagged score with a missing value dummy for the other) and demographics, classroom-level student demographics, teacher degree level, and grade and test type effects. Of the full sample of 210 teachers, 206 take the same tests with at least one other teacher. Similarly, 204 and 174 teachers were enrolled in TEPs and employed in districts with at least one other teacher. All standard errors are clustered at the teacher level. TEP = teacher education program.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

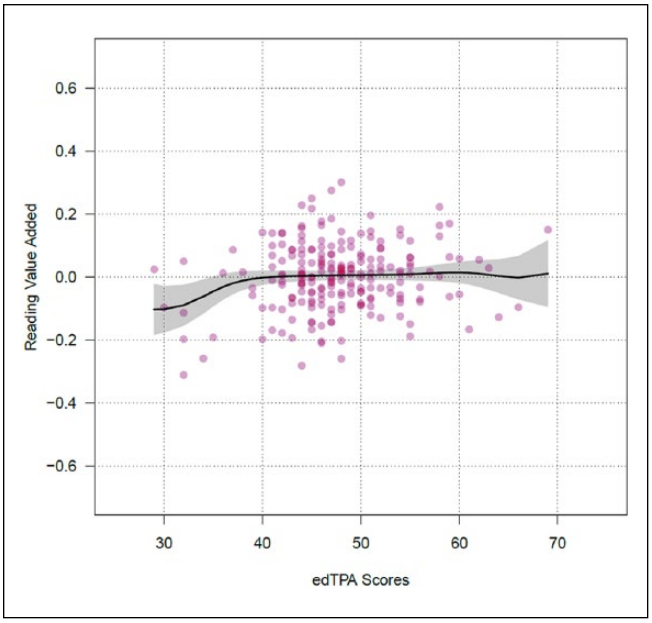

Figure 3. Relationship between edTPA scores and reading value added.

workforce the following year. Columns 7 and 8 report specifications in which the three subscores of the edTPA are separately included in the model and show that the positive relationship between the total score and the likelihood of being in the labor market is driven largely by the assessment and instruction subscores. When we consider quintiles of edTPA scores, we find that scoring in the top quintile of the edTPA is associated with a 14 percentage point increase in the probability that a candidate will be employed in the following year, as compared with a candidate who scored in the bottom quintile.

To help visualize the relationship between edTPA scores and the probability of teaching employment, Figure 2 plots the observed probability of employment associated with each edTPA score, along with a polynomial best-fit line.[29] Two patterns are worth noting. First, the relationship between edTPA scores and probability of employment is relatively steep and linear in the lower range of edTPA scores—with no discontinuity at the current passing score of 35—suggesting that, at least at the lower end of the distribution, continuous edTPA scores reflect some candidate trait or traits that are predictive of employment. Second, the relationship is much weaker in the upper range of the distribution of edTPA scores, which means that the probabilities of employment are similar for candidates within the range of relatively high edTPA scores.

Although the results in Table 3 and Figure 2 demonstrate a strong relationship between edTPA scores and the probability that a teacher candidate is employed in Washington State's K-12 public teaching workforce, it is not possible to disentangle preferences of teacher candidates and employers in interpreting these findings. As noted above, some districts may use edTPA to help them decide among teacher applicants. On the teacher candidate side, moreover, these findings may reflect the fact that more dedicated teacher candidates perform better on the assessment and are also more likely to enter the profession.

Table 5. Value-Added Results in Math (Stacked Model).

	edTPA as a screen						edTPA as a signal					
Variables of interest	1	2	3	4	5	6	7	8	9	10	11	12
Passing in Washington	0.038 (0.071)	0.061 (0.068)	0.061 (0.058)									
Future Washington passing score				0.052 (0.045)	0.085† (0.043)	0.036 (0.037)						
Total score							0.029† (0.015)	0.035* (0.016)	0.015 (0.014)			
Assessment factor										−0.004 (0.026)	0.003 (0.026)	0.016 (0.019)
Planning factor										0.060* (0.025)	0.071† (0.025)	0.002 (0.021)
Instruction factor										−0.027 (0.021)	−0.041† (0.021)	−0.001 (0.022)
TEP effects		X			X			X			X	
District effects			X			X			X			X
Teachers	206	206	206	206	206	206	206	206	206	206	206	206

Note. All models control for student prior performance (either both or just lagged or twice lagged score with a missing value dummy for the other) and demographics, classroom-level student demographics, teacher degree level, and grade and test type effects. Of the full sample of 206 teachers, 202 teachers take the same tests with at least one other teacher. Similarly, 201 and 176 teachers were enrolled in TEPs and employed in districts with at least one other teacher. All standard errors are clustered at the teacher level. TEP = teacher education program.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .00$.

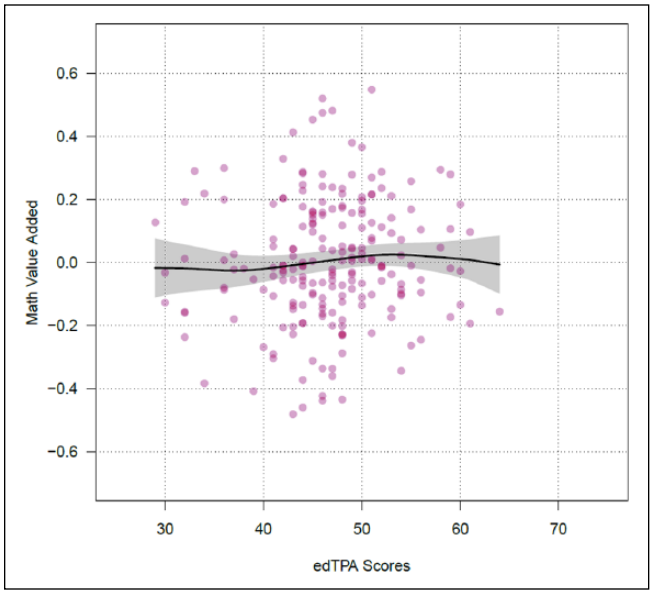

Figure 4. Relationship between edTPA scores and mathematics value added.

edTPA as a Screening Mechanism

Columns 1 to 6 of Tables 4 and 5 summarize the relationship between passing the edTPA (either at the current Washington State cut score of 35 or future cut score of 40) and teacher effectiveness in reading or mathematics, respectively. We estimate these screening models using data from the classrooms of teachers employed in the year following their edTPA administration. Given that many teacher candidates do not find teaching positions and that only a minority of teachers work in tested grades and subjects, this is a necessarily small subset of the total number of teacher candidates sitting for the edTPA. We may therefore worry that such selection biases our results. The concern is that teachers who perform poorly on the edTPA but still obtain teaching positions likely have other skills that are valued in the workplace but are not observed in our data, suggesting that the coefficients reflecting the relationship between edTPA performance and teacher effectiveness are biased downward, that is, a lower bound on the true relationship. As discussed in the previous section, there are good reasons to believe that sample selection bias is a minimal concern, but this motivates the bounding exercise described later in this section.

The models in Tables 4 and 5 correspond to Equation 2 and include lagged test scores and other student background controls (the specific independent variables used in each model specification are reported in notes below the table), but they exclude other teacher candidate variables as we are focused only in assessing the pass/fail screening value of the edTPA assessment. However, the coefficients in Tables 4 and 5 change very little when the models include additional teacher controls (such as WEST-B scores). We also note that results are very consistent between the primary specifications reported in Tables 4 and 5 and the more conservative specifications that either only use students with nonmissing prior year test scores or nonmissing twice-lagged test scores.[30]

Column 1 of Table 4 demonstrates that teacher candidates who pass the edTPA at the Washington State cut score are

more effective in reading instruction, all else equal, than teacher candidates who fail the edTPA on their first test administration. Specifically, students assigned to a teacher who passed the edTPA score 0.252 standard deviations higher, all else equal, than students who failed the edTPA. This relationship is large and statistically significant in all specifications—that is, comparing candidates to other candidates from the same TEP (column 2) or who teach in the same school district (column 3)—and are more modest but still statistically significant when we consider whether candidates would have passed the test at the future Washington State cut score. We interpret these results as suggesting that the edTPA has strong predictive validity in reading as a screen at these cut points. Our point estimates for the edTPA screening effect in mathematics in columns 1 to 6 of Table 5, however, are smaller and generally statistically insignificant. Although positive in all specifications, the screening coefficient in mathematics is statistically significant in only one specification (column 5).[31]

The differences between the screening coefficients in reading and the corresponding coefficients in math are statistically significant, and these differences are reflected in Figures 3 and 4, which plot estimated teacher value added and edTPA test scores for all teachers in our sample. The lines plotted in these figures show local linear estimates of the relationship between teacher value added and edTPA test scores.[32] While these figures do *not* control for candidate test type (and thus candidates are being compared with all other candidates regardless of test type), they illustrate that candidates who fail the edTPA at the current Washington State cutoff (35) and future Washington State cutoff (40) tend to be considerably less effective in reading (Figure 3), but less so in mathematics (Figure 4). The predicted effectiveness in reading increases sharply before the cut points, but predicted effectiveness in mathematics changes relatively little in this same range. As demonstrated by the scatter plot, we observe a smaller number of teachers with failing scores in the reading sample than in the mathematics sample and these teachers are more likely to have low value added.[33]

The Signal Value of edTPA Performance

The value of the edTPA as a signal of teacher quality is an important policy issue. Recall that the edTPA is described as an "educative assessment," and this is much more plausible if there is predictive validity to the assessment away from the cut point (suggesting that changes in performance by candidates or institutions are indeed predictive of teacher effectiveness). In addition, whether inservice teachers with higher edTPA scores are more effective is an important policy question given that school systems may wish to consider an applicant's edTPA scores in making hiring decisions.

Columns 7 to 12 of Tables 4 and 5 report the estimated relationships between continuous measures of candidate edTPA performance and student achievement in reading and mathematics, respectively. Columns 7 to 9 of Table 4 illustrate that we find little evidence that edTPA scores throughout the distribution are predictive of teacher effectiveness in reading. Specifically, the coefficient in column 7 means that a one standard deviation increase in a candidate's edTPA score is correlated with a 0.02 standard deviation increase in student performance in the candidate's classroom in his or her first-year teaching, but this relationship is not statistically significant. The weak relationship between continuous edTPA scores and teacher effectiveness in reading is reflected in Figure 3, as there is little increase in predicted teacher effectiveness within the range of passing scores (i.e., above a 40). We note, however, that this relationship is positive and statistically significant when we focus solely on candidates who took the edTPA after it became consequential in January 2014.[34]

However, columns 7 to 12 of Table 5 provide some evidence that edTPA scores provide a signal of future teacher effectiveness in mathematics.[35] Specifically, when candidates are compared across TEPs and districts (column 7), a one standard deviation increase in a candidate's edTPA score is correlated with a 0.03 standard deviation increase in student performance in the candidate's classroom in his or her first-year teaching, and this relationship is marginally statistically significant. This is reflected in the generally positive slope of the local linear fit line in Figure 4.

The relationship between edTPA scores and mathematics teaching effectiveness is stronger when candidates are compared with other candidates from the same TEP (column 8), but weaker when candidates are compared with other candidates who are teaching in the same school district (column 9). As discussed in Goldhaber et al. (2013), it is possible that the district fixed effects in the model in column 9 capture district-level effects that are attributed to teachers in the estimates reported in columns 7 and 8, but it is also possible that these effects remove average differences in teacher quality among different school districts that should be attributed to teachers. Given that we cannot distinguish between these possibilities, we simply conclude that the predictive validity of the edTPA as a signal of future teaching effectiveness in mathematics is stronger when comparisons are made across districts rather than within districts.

Finally, columns 9 to 12 of Table 5 consider the three edTPA subscores as joint predictors of teacher effectiveness in mathematics, and suggest that candidate performance on the Planning rubrics are driving the relationships in columns 7 to 9. This is an interesting finding, as the Planning subscores were less predictive of the probability of employment than were the other two subscores (see Table 3).

Policy Implications

In this study, we find that teachers failing the edTPA under the future Washington State passing threshold have lower value added in reading than teachers who passed the test at

Table 6. Conditional Probabilities of Teacher Effectiveness Given edTPA Performance.

	Stacked math sample		Stacked reading sample	
Quintile of value added	Fail	Pass	Fail	Pass
Bottom quintile	0.190 (0.088)	0.202 (0.029)	0.462** (0.110)	0.185 (0.028)
Top quintile	0.143 (0.087)	0.202 (0.030)	0.077 (0.111)	0.205 (0.029)

Note. Each cell gives the probability that a teacher with the indicated performance on the edTPA falls into each quintile of the value-added distribution. Standard errors in parentheses. The test of significance is against the null hypothesis that the proportion is 0.2.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

this cut score. We find no statistically significant difference between those who pass and those who fail in mathematics, although changes in the assessment score are predictive of teacher performance. These results generally hold when a licensure test of candidates' basic skills is included in the model, which suggests that portfolio-based assessments such as the edTPA contain information about teaching practice that is not captured by these basic-skills tests. Although our point estimates are imprecisely estimated due to the small samples employed in this study, the magnitudes of the signal estimates are roughly similar to those observed in studies of other licensure tests (Clotfelter, Ladd, & Vigdor, 2007; Darling-Hammond et al., 2013; Goldhaber, 2007).

To put the results in perspective, we estimate the probability that a teacher candidate failing the edTPA is a *low-performing teacher* (defined as being in the bottom 20% of value added) or a *high-performing teacher* (defined as being in the top 20% of value added).[36] The results of this test are in Table 6. If passing the edTPA provided no predictive power for value added, we would expect 20% of teachers who fail the test to be in each of these categories. Not surprisingly, given the null screening results in mathematics, we find that 19% of mathematics teachers who fail the edTPA are in the low-performing category. On the contrary, we find that 46% of reading teachers who fail the edTPA are in the low-performing category, far higher than the 20% we would expect by chance alone. That said, if the edTPA really were used as a one-time, high-stakes test for employment eligibility, screening these candidates who would become ineffective teachers comes at the cost of screening out some candidates who would become effective teachers. Specifically, 8% of reading teachers and 14% of math teachers who fail the edTPA are in the high-performing category (top 20% of value added); neither of these proportions is statistically different than the 20% we would expect by chance.

We can more simply summarize these proportions using the "number needed to treat." In medicine, the number needed to treat is the average number of patients that would need to be assigned an intervention to avoid one additional adverse outcome. A low number needed to treat indicates an efficient intervention as it implies that a greater number of patients benefit. In this case, we can identify the number of test takers needed to screen out a lowest quintile teacher. We do a back-of-the-envelope calculation suggesting that the edTPA identifies one bottom quintile reading teacher for every 17 assessed candidates, while it identifies one bottom quintile mathematics teacher for every 39 candidates. Put another way, this suggests a cost in exam fees to candidates of US$5,100 to identify an ineffective reading teacher and US$11,700 to identify an ineffective mathematics teacher.

Conclusion

Given that this is the first predictive validity study of the edTPA, and given the nuanced findings we describe above, we are hesitant to draw broad conclusions about the extent to which edTPA implementation will improve the quality of the teacher workforce. Instead, we relate our findings back to the different theories of action for how the edTPA *might* improve teacher workforce quality, but we stress that even these conclusions come with important caveats and trade-offs that policy makers and teacher educators should weigh as they interpret these results.

The first theory of action is that the edTPA can be used as a screen to prevent ineffective teacher candidates from entering the workforce. The screening results in reading—demonstrating predictive validity around the current and future Washington State cut points used for licensing decisions—generally suggest that this theory of action is promising in terms of improving overall workforce quality in reading. But as we discuss in the previous section, this screening comes at a cost, as candidates who fail the edTPA but become high-performing teachers will also be screened out of the workforce. We do not find evidence of a screening effect in mathematics, although our estimates are imprecisely estimated. This relationship may, in part, be caused by the edTPA's focus on candidates' writing capacities, which may be more related to a teacher's ability to teach reading than mathematics.[37] It is also important to recognize that the screening theory of action is predicated on teacher candidates failing the assessment. It is unclear that this screening theory of action can actually work in a setting in which candidates are able to take the test multiple times to pass, as the ability of the assessment to predict teacher effectiveness is likely to be low for candidates with multiple retakes (Cowan & Goldhaber, 2016).

The second theory of action is that the edTPA could improve the quality of *all* teaching candidates through the

experience of the assessment or programmatic changes that are related to information TEPs receive about teacher candidate performance. This is much more likely if the edTPA scores can serve as a signal of quality teaching beyond just at the cut point required to participate in the labor market. In this case, it is the modest but statistically significant results in mathematics that suggest promise for this theory of action and the weaker results in reading that suggest caution. That said, the extent to which the edTPA can "support candidate learning and preparation program renewal" (edTPA, 2015) likely depends on the ability of TEPs to create feedback loops that allow candidate performance on the edTPA to influence the training they provide. Moreover, policy makers and teacher educators also need to weigh these results against the possibility that the high-stakes use of the edTPA may adversely affect the diversity of the teacher workforce, given the large differences between the passing rates of White and Hispanic teacher candidates in Washington.

We believe there are a number of potential next steps that are not possible to pursue with the data used in this study but that would be valuable to policy makers and teacher educators. One is to investigate the degree to which the different rubric scores within the edTPA might be reweighted (or modified) to increase the relationship between summative edTPA scores and student achievement or teacher value added. The samples in Washington State are currently insufficient for optimal weighting exercises (e.g., Goldhaber, Grout, & Huntington-Klein, 2014), but such exercises are possible with additional years of data and/or data from other states and would be valuable to TEPs looking to prioritize different aspects of their training of teacher candidates. A second next step might be to assess how edTPA scores are related to other, broader measures of teacher performance, such as observational ratings. This is not currently possible using Washington State's administrative data, but it may be possible elsewhere. Finally, given concerns about the fairness of teacher observations across classroom contexts (Steinberg & Garrett, 2016) and recent calls to place more student teachers in disadvantaged schools (Krieg, Theobald, & Goldhaber, 2016), policy makers would benefit from evidence about whether edTPA scores vary substantially across teacher candidates in different kinds of student teaching positions.

A final caveat to these conclusions—and an essential issue for policy makers and teacher educators to weigh in interpreting these results—is whether the results we reference above justify the investments that candidates, states, and TEPs have made in the edTPA. While the monetary costs associated with the edTPA are easily quantifiable (e.g., US$300 per teacher candidate), there are also less easily quantifiable time-commitment costs for both candidates and programs. We know very little regarding whether these costs might affect the pool of people who seek to become teachers. We therefore view the interpretation of these results as very much in the eye of the beholder, and we hope this early analysis spurs an evidence-based discussion about the potential promise and drawbacks of edTPA implementation.

Authors' Note

The views expressed in this article do not necessarily reflect those of American Institutes for Research. Responsibility for any and all errors rests solely with the authors.

Acknowledgments

The authors thank Trevor Gratz for outstanding research assistance and Joe Koski, John Krieg, Gerhard Ottehenning, Ray Pecheone, Amy Vaughn, Jennifer Wallace, and Andrea Whitaker for helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER; IES Grant R305C120008), the Bill and Melinda Gates Foundation (Grant OPP1128040), and an anonymous foundation.

Notes

1. See http://edtpa.aacte.org
2. For a full summary of edTPA participation across the country, see edTPA (2015), p. 13.
3. Note that the existence of different cut points in different states means that the edTPA cannot be expected to have predictive validity "only" around a single cut point.
4. All analytic models presented in this article control for test type, so compare outcomes only between candidates who took the same test type.
5. Candidates may receive an incomplete score on any of the 15 rubrics for having technical issues with the upload, uploading an incomplete file, having an edited video, or uploading material that is not related to the handbook. If a candidate received only one incomplete score, it counts as a zero in the calculation of the final summative score; but the summative score is incomplete if the candidate receives an incomplete on two or more rubrics.
6. We performed a principal components analysis on the 15 rubric scores and found that the rubric scores load onto three factors that align closely with these areas.
7. The national edTPA handbook for elementary education also includes three additional mathematics assessment rubrics (e.g., "Analyzing Whole Class Misunderstandings") that have not been adopted in Washington State.
8. See http://edtpa.aacte.org/about-edtpa#Goals-1
9. The 2014 edTPA administrative report states that "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" (edTPA, 2015, p. 24). However, these data have never been published,

and follow-up documentation from the authors suggests that these relationships are more mixed than this quote suggests (Susan Benner, personal communication, May 2016).

10. Darling-Hammond et al. (2013) report nearly identical point estimates as those reported in this article but with substantially more precision using a considerably smaller sample than is available in this article. We attempted to replicate their findings using differing assumptions regarding the appropriate level of clustering and could only estimate coefficients with similar levels of precision in models that assume independent errors across students in the same classroom. We attempted to compare modeling choices directly, but in discussions with the authors, we were unable to do so as they no longer have their data files (Linda Darling-Hammond, personal communication, February 2016).
11. See http://www.ctc.ca.gov/commission/agendas/2012-09/2012-09-2F.pdf
12. See http://assessment.pesb.wa.gov/performance-assessments/important-links-edtpa-information/edtpa-policies
13. The correlations between the three subscores range from .598 to .661.
14. We drop incomplete scores in cases where the candidate resubmits materials within a month of the score reporting date. We experimented with models that consider all incomplete scores as failures and found similar results.
15. Some alternative licensing exams may be submitted instead of taking the Washington Educator Skills Test–Basic (WEST-B). Thus, not all prospective teachers take the WEST-B (RCW 28A.410.220 & WAC 181-01-002).
16. About one third of Washington State schools participated in the state's Smarter Balanced Assessment pilot in the 2013-2014 school year, so test scores are not available in 2013-2014 for students in these schools. We discuss our approach to these missing data in the analytic approach section.
17. Comprehensive Education Data and Research System (CEDARS) data include fields designed to link students to their individual teachers, based on reported schedules. However, limitations of reporting standards and practices across the state may result in ambiguities or inaccuracies around these links. We limit the student sample to students who received instruction from a single teacher in that subject and year.
18. Note that the quintiles in this table are based on edTPA scores across multiple test types, but all models include fixed effects for test type (so candidates are compared only with other candidates who took the same test type).
19. The correlations between continuous edTPA scores and the three WEST-B subtests are moderate (r = .20 in mathematics and reading, r = .25 in writing).
20. This is consistent with research showing that performance on licensure tests varies across teacher candidate subgroups (Goldhaber & Hansen, 2010).
21. These results are robust to controlling for candidate TEP (that is, Hispanic candidates are more likely to fail the edTPA than non-Hispanic White candidates within the same TEP) and conflict with recent evidence (edTPA, 2016) from a national census of edTPA test takers that finds Black teacher candidate scores to be significantly lower than the scores of White candidates, but no significant difference between White and Hispanic teacher candidate edTPA performance.
22. Hispanic teacher candidates are also considerably more likely than White candidates to score lower than a 40 (the state's future cut score), though we can not necessarily conclude that this difference in passing rates would hold under this new cut score.
23. As discussed in "Assessment of Prospective Teachers and the Role of the edTPA" section, there are 27 different versions of the edTPA, so this ensures that candidates are only compared with other candidates who completed the same test type.
24. Note that we do not need to control for teaching experience because every teacher in the value-added model (VAM) sample is a first-year teacher.
25. We also experiment with school fixed effects models, but a relatively small number of teachers in the VAM sample teach in the same school as compared with other teachers who took the edTPA.
26. These results are provided in Tables A2 to A5 in the online appendix.
27. The coefficients we discuss are not reported in the tables but are available from the authors upon request.
28. These results are reported in Table A1 in the online appendix.
29. The best-fit line is estimated from a logit at the teacher candidate level, with the order of polynomial chosen to minimize the Akaike information criterion (AIC) of the regression.
30. These results are reported in Tables A2 to A5 in the online appendix.
31. As shown in Tables A6 and A7 in the online appendix, the screening results are similar when we estimate models only for candidates who took the edTPA after it became consequential. The differences between the screening results before and after the edTPA became consequential are not statistically significant.
32. We estimate teacher value added using the same specification as Equation 2, but omitting the edTPA scores and teacher controls. We then estimate local linear regressions of estimated teacher value added on edTPA scores using the np package in R (Hayfield & Racine, 2008).
33. To obtain an estimate of the potential magnitude of sample selection bias in these estimates, we conduct a bounding exercise in the spirit of Lee (2009). Our results suggest the point estimates for the screening effect lie between 0.05 and 0.40 for reading and between −0.09 and 0.09 in mathematics. Results are available from the authors upon request.
34. See Table A7 in the online appendix. The difference between this relationship before and after the edTPA became consequential is not statistically significant.
35. Note, however, that the differences between the signal coefficients in math and the corresponding coefficients in reading are not statistically significant.
36. We obtain similar results if we instead estimate these conditional probabilities using the simulation method suggested by Jacob and Lefgren (2008).
37. For example, edTPA scores are more highly correlated with WEST-B writing scores (r = .25) than WEST-B reading or mathematics scores (r = .20).

References

Angrist, J. D., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, *27*(5), 483-503.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416-440.

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2008). *National board certification and teacher effectiveness: Evidence from a random assignment experiment* (No. w14608). Cambridge, MA: National Bureau of Economic Research.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593-2632.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Cowan, J., & Goldhaber, D. (2014). *Assessing the relationship between teacher performance on Washington State's ProTeach Portfolio and Student Test Performance* (CEDR Working Paper 2014-2). Seattle, WA: University of Washington.

Cowan, J., & Goldhaber, D. (2016). National Board Certification and teacher effectiveness: Evidence from Washington State. *Journal of Research on Educational Effectiveness*, *9*, 233-258.

Darling-Hammond, L. (2009). Teaching and the change wars: The professionalism hypothesis. In A. Hargreaves & M. Fullan (Eds.), *Change wars* (pp. 45-68). Bloomington, IN: Solution Tree.

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

edTPA. (2015, September). *Educative assessment & meaningful support: 2014 edTPA administrative report*. Retrieved from https://secure.aacte.org/apps/rl/res_get.php?fid=2183&ref=edtpa

edTPA. (2016, October). *Educative assessment & meaningful support: 2015 edTPA administrative report*. Retrieved from https://secure.aacte.org/apps/rl/res_get.php?fid=3013&ref=edtpa

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources*, *42*(4), 765-794.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics*, *89*(1), 134-150.

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2014, December). *Screen twice, cut once: Assessing the predictive validity of teacher selection tools* (CEDR Working Paper No. 2014-9). Seattle, WA: Center for Education Data and Research.

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.

Goldhaber, D., Krieg, J., & Theobald, R. (2014). Knocking on the door to the teaching profession? Modeling the entry of prospective teachers into the workforce. *Economics of Education Review*, *42*, 106-124.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Evaluating teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29-44.

Gottlieb, J., Hutt, E., & Cohen, J. (2016, March). *Diffusion in a vacuum: The case of edTPA*. AEFP 2016 Conference Paper, Denver, CO.

Greenblatt, D., & O'Hara, K. (2015, Summer). Buyer beware: Lessons learned from edTPA implementation in New York State. *The NEA Higher Education Journal*, 57-68.

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, *40*, 183-204.

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*(5), 1-32.

Hill, D., Hansen, D., & Stumbo, C. (2011, April). *Policy considerations for states participating in the Teacher Performance Assessment Consortium (TPAC)*. Washington, DC: Council of Chief State School Officers.

Jacob, B., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2016). *Teacher applicant hiring and teacher performance: Evidence from DC public schools* (No. 22054). Cambridge, MA: National Bureau of Economic Research.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101-136.

Jette, A. (2014, July 19). 10 tips for edTPA success. *Education Week Teacher*. Retrieved from http://www.edweek.org/tm/articles/2014/07/29/ctq_jette_edtpa.html

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Measures of Effective Teaching (MET) Project. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://eric.ed.gov/?id=ED540959

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180-195.

Krieg, J., Theobald, R., & Goldhaber, D. (2016). A foot in the door: Exploring the role of student teaching assignments in teachers' initial job placements. *Educational Evaluation and Policy Analysis*, *38*, 364-388.

Larsen, B. (2015). *Occupational licensing and quality: Distributional and heterogeneous effects in the teaching profession*. Retrieved from http://web.stanford.edu/~bjlarsen/Larsen%20(2015)%20Occupational%20licensing%20and%20quality.pdf

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*(3), 1071-1102.

Mihaly, K., McCaffrey, D., Sass, T. R., & Lockwood, J. R. (2013). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy*, *8*(4), 459-493.

Newton, S. P. (2010). *Preservice performance assessment and teacher early career effectiveness: Preliminary findings on the Performance Assessment for California Teachers*. Stanford, CA: Stanford Center for Assessment, Learning, and Equity.

Pecheone, R., Shear, B., Whittaker, A., & Darling-Hammond, L. (2013, November). *2013 edTPA field test: Summary report*.

Stanford, CA: Stanford Center for Assessment, Learning, and Equity.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, *6*(1), 43-74.

Steinberg, M., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*, 293-317.

U.S. Department of Education. (2009). *U.S. Secretary of Education Arne Duncan says colleges of education must improve for reforms to succeed*. Retrieved from http://www.ed.gov/news/press-releases/us-secretary-education-arne-duncan-says-colleges-education-must-improve-reforms-succeed

Wiswall, M. (2007). *Licensing and occupational sorting in the market for teachers*. Retrieved from http://www.econ.nyu.edu/user/wiswall/research/wiswall_teacher_licensing.pdf

Author Biographies

Dan Goldhaber serves as the director of National Center for Analysis of Longitudinal Data in Education Research (CALDER) at American Institutes for Research and the director of the Center for Education Data & Research (CEDR) at the University of Washington Bothell. His work focuses on issues of educational productivity and reform at the K-12 level, the broad array of human capital policies that influence the composition, distribution, and quality of teachers, and connections between students' K-12 experiences and postsecondary outcomes.

James Cowan is a researcher in CALDER at American Institutes for Research. His research focuses on education policy and teacher labor markets.

Roddy Theobald is a researcher in CALDER at American Institutes for Research. His research interests are in teacher education, teacher evaluation, special education, and teacher collective bargaining.