NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

AIR

State Ratings of Educator Preparation Programs: Connecting Program Review to Teacher Effectiveness

Meagan Comb
James Cowan
Dan Goldhaber
Zeyu Jin
Roddy Theobald

WORKING PAPER No. 249-0321 • March 2021

State Ratings of Educator Preparation Programs: Connecting Program Review to Teacher Effectiveness

Meagan Comb
Boston University

James Cowan
CALDER, American Institutes for Research

Dan Goldhaber
CALDER, American Institutes for Research
University of Washington

Zeyu Jin
CALDER, American Institutes for Research

Roddy Theobald
CALDER, American Institutes for Research

Contents

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H170025 to the American Institutes for Research (AIR). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors wish to thank partners at the Massachusetts Department of Elementary and Secondary Education, including Claire Abbott, Matt Deninger, Bob Lee, Liz Losee, Elana McDermott, Adrienne Murphy, Shelagh Peoples, Heather Peske, Sandra Sarucia, and Aubree Webb, for comments that improved this analysis. Finally, the authors want to note that author Meagan Comb was involved in the development and implementation of the program review process.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

State Ratings of Educator Preparation Programs: Connecting Program Review to Teacher Effectiveness

Meagan Comb, James Cowan, Dan Goldhaber, Zeyu Jin, Roddy Theobald
CALDER Working Paper No. 249-0321
March 2021

Abstract

States are responsible for setting and evaluating the standards that teacher preparation programs (TPPs) must meet for accreditation. Despite the considerable investment that states make in this process, no prior research has linked the ratings of TPPs generated by program reviews to inservice teacher performance. In this paper, we describe analyses of program review ratings from Massachusetts and their relationship to formal inservice teacher evaluation ratings and the value-added effectiveness of teachers. When comparisons are made across all schools and districts in the state, we find that a TPP's review scores are positively predictive of both inservice teacher evaluations and value added of TPP graduates, particularly when scores are aggregated within specific categories like partnerships and field-based practices. These relationships, however, become more modest for teacher evaluations and statistically insignificant for value added when the relationships are identified based on comparisons between TPP graduates who are teaching in the same schools and districts. It is not possible to separate whether these differences are due to the TPPs, the schools and districts themselves, or the connections between them, so future work is necessary to further validate TPP review scores in this setting and others.

1. Introduction

The role that teacher preparation providers (TPPs) play in shaping the teacher workforce has been researched extensively and has received a great deal of policy attention. This makes sense given that improving teacher quality appears to be an important means of affecting student achievement,[1] and TPPs are thought to play a pivotal role in influencing the quality of new teacher entrants to the labor market (Goldhaber, 2019). Yet despite a decade of national focus on what constitutes effective teacher preparation, our specific knowledge in this regard is limited.[2]

States are responsible for setting the standards that programs must meet in order to grant prospective teachers with a credential necessary to qualify to be a teacher. Thus, states have enormous influence and opportunity to shape the experiences that teacher candidates have in their preservice training before they enter the workforce through accreditation and program approval requirements (CCSSO, 2012; Feuer, Floden, Chudowsky, & Ahn, 2013).[3] But while considerable research attributes teacher effectiveness to the specific TPP from which teachers graduated (e.g., Bastian, Lys, & Pan, 2018; Boyd et al., 2009; Goldhaber, Liddle, & Theobald, 2013; Koedel, Parsons, Podgursky, & Ehlert, 2015; Mihaly, McCaffrey, Sass, & Lockwood, 2013; Ronfeldt & Campbell, 2016; von Hippel et al., 2016), program review processes like the one used in Massachusetts provide an opportunity to understand more complex and dynamic features of teacher preparation and their relationships to graduate outcomes.

[1] Teachers are arguably the most important schooling input that influences student outcomes (e.g., see Chetty, Friedman, & Rockoff, 2014; Goldhaber, Brewer, & Anderson, 1999; Jackson, 2018; Kraft, 2019; Hanushek & Rivkin, 2010; Nye, Konstantopoulos, & Hedges, 2004).

[2] For instance, a new report by the National Academy of Sciences concludes that we know relatively little empirically about the criteria for admission or curricular requirements that lead to teacher candidates who leave TPPs with a skillset acceptable for first-year teachers (National Academy of Sciences, 2020). For other consensus reports on what we know about teacher preparation, see, for instance, Feuer and colleagues (2013), Holmes Group (1986), and National Commission on Teaching and America's Future (1996).

[3] Some states require that TPPs be accredited by national accrediting bodies, such as the Council for the Accreditation of Educator Preparation (CAEP), as a requirement for state approval. Prospective teachers must also pass various licensure tests to be fully eligible to teach (Goldhaber, 2007).

In this paper, we describe research that connects the ratings of TPPs collected during the comprehensive program review process in Massachusetts to both inservice teacher evaluation ratings and value added of TPP graduates. We find that a TPP's review scores are positively and significantly predictive of both inservice teacher evaluation ratings and value added when comparisons are made across all schools and districts in the state. These relationships are robust to controlling for differences in licensure test scores across TPPs, but they are not robust to comparisons of TPP graduates who are teaching in the same schools and districts. This implies that schools and districts with higher teacher evaluation ratings and test-score gains tend to employ more teachers from TPPs with higher program review scores. Unfortunately, we cannot definitively determine whether this is reflective of true differences in the performance of teachers who sort into particular schools and districts, or an artifact of the performance evaluations or value added of the teachers who end up in those schools and districts.

The remainder of the paper proceeds as follows. In Section 2, we provide some background about prior research regarding teacher preparation requirements and the specifics of the program approval process in Massachusetts. We describe our data and measures in Section 3, outline our analytic approach in Section 4, and present the results in Section 5. Finally, in Section 6 we discuss the policy implications of this study as well as areas for further research.

2. Background

2.1 TPP Review and TPP Research in Other States

Teacher preparation programs (TPPs) must obtain state approval in order to be eligible to prepare teacher candidates for employment in K–12 public schools, and the approval processes vary across states (Zeichner, 2006). However, relatively little systemic work connects program

requirements to inservice teacher and student outcomes (Goldhaber, 2019; Ronfeldt & Campbell, 2016). Yet despite the lack of rigorous evidence on the extent to which TPPs contribute to the development of teacher candidates or produce graduates with teaching skills deemed to be adequate for first-year teachers, there are long-standing critiques of TPPs and the accreditation processes under which they operate. A prominent example is that of Arthur Levine (2006), former president of Teachers College, Columbia University, who noted that "[u]nder the existing system of quality control, too many weak programs have achieved state approval and been granted accreditation" (p. 61).[4]

TPPs are generally evaluated based on subjective judgments as part of a program review.[5] In Section 2.2, we discuss the program review specifically in Massachusetts but first provide an overview of a typical program review process in states across the country. The specific criteria upon which TPPs are judged and the means by which they are judged differ somewhat from state to state, but many of these efforts evaluate some combination of the following: entrance requirements; the number of candidates being prepared in high-need subjects; curricular requirements; course syllabi and textbooks; faculty qualifications (e.g., part-time and full-time degree levels); and fieldwork requirements, such as hours of clinical practice (Feuer et al., 2013). States may also require that TPPs be accredited using standards developed by one of two national accrediting bodies: the Council for the Accreditation of Educator Preparation (CAEP) or the more recently formed Association for Advancing Quality in Educator Preparation (AAQEP).[6] Both state agencies and TPPs invest significant time and resources in

[4] See also, for instance, criticisms by former U.S. Secretary of Education Arne Duncan (n.d.) or, more recently, the National Council on Teacher Quality (Greenberg, Pomerance, & Walsh, 2011).
[5] There had been a brief federal push to use criteria such as entry rates in the public school teaching workforce and the test achievement of teachers to help inform judgments about TPPs (von Hippel & Bellows, 2018).
[6] See http://caepnet.org/standards/2022/introduction and https://aaqep.org/wp-content/uploads/2020/01/2020-Guide-to-AAQEP-Accreditation.pdf. Massachusetts does not have a formal partnership with either organization.

these efforts as a seemingly impactful quality monitoring and improvement endeavor, yet there are no published studies that speak to their efficacy, or even whether the judgments about TPPs are related to the performance of the teacher candidates who end up employed in public schools.

Judgments about TPPs inform high-stakes decisions such as program accreditation. Therefore, these judgements are the centerpiece of state influence on the preparation of new teachers. But, as summarized in Feuer and colleagues (2013): "Most measures of [TPP] quality in use today seem to have been chosen based on their face validity—in other words, they appear to address important characteristics of teachers and teaching—and on the feasibility of collecting the data, rather than on empirical correlations or 'predictive validity' evidence linking qualities of teacher preparation with student outcomes." (p. 26). Indeed, we were unable to find a single published study that directly assesses whether ratings of TPPs by states or outside accreditation agencies, connected to program reviews, are correlated with teacher or student outcomes.

Despite periods during which the federal government appeared likely to take an active role in TPP accountability, the states primarily determine whether TPPs can prepare prospective teachers.[7] There is, however, little consensus about precisely how TPPs ought to be evaluated, despite research in the last decade connecting various outcome measures back to TPPs. For instance, a number of studies have aggregated the value added of teachers in order to obtain TPP-based value-added measures (e.g., Boyd et al., 2009; Goldhaber et al., 2013; Koedel et al., 2015; Mihaly et al., 2013; von Hippel et al., 2016). While these studies reach somewhat different

[7] Under the Obama administration, the U.S. Department of Education signaled that understanding the effectiveness of teacher preparation was a national priority through regulations in Title II of the Higher Education Act that called for states and preparation organizations to collect data and publicly report on placement and retention of graduates in teaching positions, feedback from administrators about the competence of graduates, and the effectiveness of graduates in raising student achievement (U.S. Department of Education, 2014). These regulations were subsequently rescinded under the Trump administration, though a number of states do require such reporting. "Public accountability" is also a feature of the TPP landscape with non-governmental organizations, like the National Council on Teacher Quality (NCTQ) ratings of TPPs (Goldhaber & Koedel, 2019).

conclusions about whether TPPs explain an educationally meaningful proportion of the variation in student achievement, all find that there is substantially more variation in teacher effectiveness within programs than between them. In their review of TPP research across a number of states, von Hippel and Bellows (2018) conclude that "teacher quality differences between most [TPPs] are negligible—0.01–0.03 standard deviations of student test scores—even in states where larger differences were reported previously" (p. 296).[8] But importantly for the purposes of this study, prior research on the variation in teacher effectiveness of TPP graduates *in Massachusetts* (Cowan, Goldhaber, & Theobald, 2017) finds that the variation in teacher value added explained by individual TPPs in Massachusetts is on the high end of what has been found in other states; the estimates for the least effective TPPs relative to the state average correspond to about 5 to 10 weeks of student learning in math and about 6 to 20 weeks of learning in English language arts (ELA).

There are fewer studies that focus on non-test outcomes (Bastian et al., 2018; Ronfeldt & Campell, 2016), but these tend to find larger effects associated with the teachers aggregated to the TPP level. Ronfeldt and Campbell (2016), for instance, analyze the association between teacher observational ratings and TPPs in Tennessee and suggest that "…using observational ratings to evaluate [TPPs] has promise [as they] were able to detect significant and meaningful differences between TPPs, which were fairly robust across modeling approaches, [and the TPP] rankings based on observational ratings were positively and significantly related to rankings

[8] The von Hippel and Bellows conclusion is based on a statistician's standard of evidence about what constitutes differences between TPPs; it is not clear that this is the right standard for judging TPPs given that other methods of evaluating programs (discussed below) are also subject to errors that lead to uncertainty about estimated differences between programs (Conaway & Goldhaber, 2020).

based on student achievement gains" (abstract).[9] Importantly, these studies reflect an attempt to discern TPP quality through the use of a single measure (e.g., value added, observation scores), whereas most program review processes, including the process used in Massachusetts, attempt a more comprehensive collection of evidence and outcomes around which summative judgments are made.

2.2 TPP Review in Massachusetts

In Massachusetts, the stated purpose of review and approval of TPPs is "to assure the public that educators who complete preparation programs in Massachusetts are prepared to be effective educators ready to support the success of all students in the Commonwealth" (Massachusetts Department of Elementary and Secondary Education, 2020). The framework for accomplishing this goal during the period that we study was established in 2012, when the Massachusetts Board of Elementary and Secondary Education introduced new regulatory requirements for the TPP approval process,[10] and was implemented for the first time in 2014.

The Massachusetts Department of Elementary and Secondary Education (DESE), which is responsible for TPP reviews, focuses reviews on organizational aspects of TPPs, such as their systems for continuous improvement, partnerships with school districts, and quality of clinical experiences. The examination of instructional programming is part of these expectations but not the primary focus for decision making about ongoing approval. The 2012 regulatory reform of

[9] Importantly, findings about both the value-added effectiveness and performance of teachers receiving their credentials from a particular program are not necessarily indicative of the *contribution* that TPPs make toward the development of teacher candidates. There are various types of selection—from selection of applicants into TPPs to selection of teacher candidates from particular TPPs into specific schools and districts—such that the TPP effects may not reflect the value of the educational experiences that those teachers received while attending those programs (Goldhaber & Ronfeldt, 2020).

[10] The 2012 regulatory changes emphasized partnerships between TPPs and PK–12 districts and the alignment of expectations for systems of continuous improvement; increased requirements around the clinical experience; and data collection, reporting, and accountability for the outcomes of candidates once employed. The 2012 changes also modified requirements for individual teacher candidates and their supervision, for instance, increasing required hours of student teaching. For more details on the overall process, see https://www.doe.mass.edu/edprep/review/.

program approval standards also required DESE to emphasize the outcomes of new TPP graduates who become teachers ("output-based evaluation") in the process. This includes surveys of principals about the readiness of newly hired teachers from different TPPs, information on the employment of teacher candidates, the educator evaluation ratings (the inservice evaluation ratings that teachers receive), and student growth data for TPP completers who enter the Massachusetts public school teaching workforce.[11] DESE, not the individual TPPs, assumes responsibility for actively collecting, compiling, and disseminating these outcome data.

The DESE review process focuses on evaluating the evidence that TPPs are successfully preparing teacher candidates, rather than focusing on whether particular TPP practices are "good" or not. As described in Section 2.1 above, other states and CAEP have established detailed rubrics to describe what it looks like to meet a standard, often emphasizing the inputs in place (e.g., entry requirements, hours in the field). This differs from DESE's process, which rates expectations solely on the sufficiency of evidence present to justify the expectation being met.[12]

TPP reviews are conducted by DESE with the support of four to six volunteer individuals (half from the PPK–12 sector and half from TPPs) who are intentionally selected and trained. Together, the team reviews evidence for specific criteria in five domains,[13] summarized in Table 1. Each domain encompasses between four and 12 criteria that further define the

[11] DESE recognized that TPPs have limited ability to collect and aggregate outcome measures for the teacher candidates that they graduate, so states made significant investments to ensure that outcome-based measures were available and provided back to TPPs.

[12] For instance, while many other states and CAEP have set minimum grade point average (GPA) thresholds for entrance into teacher preparation (e.g., cohorts must have an average GPA of 3.0), Massachusetts has intentionally avoided doing so, weighting more heavily the impact of whatever practices the TPP chooses to institute. Similarly, there are no embedded benchmarks for the rates or percentages for any of the outcome measures available in the state. All are considered individually within the full context of the review and specific to the TPP; see https://www.doe.mass.edu/edprep/review/evaltool/.

[13] A sixth domain, Instruction, focuses on programmatic coursework specific to the subject and grade level of the license being sought. This is the one domain in the Massachusetts TPP review process that is evaluated at the individual program level, not the overall TPP level. Performance on these instruction domains certainly influences the overall approval determinations, but the most significant weight is placed on the five organization level domains. As a result, these five organization-level domains are the foci of our analysis.

expectations against which TPP evidence is evaluated. Using judgments made about individual criteria being met, reviewers then rate each domain on a four-point scale from "Exemplary" to "Unsatisfactory." DESE and reviewers rely on calibration and consensus in these decisions. There is no set formula or calculation for how criteria ratings roll up to judgments at the domain level. The domain-level ratings (described in more detail below) are then further considered in a recommendation around an overall approval status. TPPs receive one of five approval designations: "Approved with Distinction," "Approved," "Approved with Conditions," "Probationary Approval," and "Not Approved."[14]

In the year before the enactment of the program approval reforms, there were 71 TPPs operating more than 1,700 individual teacher licensure programs (e.g., elementary or math grades 8–12 programs).[15] TPPs include institutions of higher education, nonprofit operators (e.g., Teach For America, Boston Teacher Residency), districts, professional associations, and educational collaboratives. Regardless of the type of organization operating the program, all providers undergo the same review process.

DESE formally evaluated 47 of these TPPs between 2014–15 and 2019–20.[16] TPPs were scheduled into review cohorts based on the timing since their last review and their relative size to create a balance in the lift required to conduct reviews each year. Table 2 describes the distribution of approval designations over this period. About two-thirds (31) of these TPPs received an "approved" designation, four were approved with distinction, eight were approved with conditions, and four received probationary approval. While Massachusetts has not used a

[14] See https://www.doe.mass.edu/edprep/resources/guidelines-advisories/program-approval//.
[15] These numbers have shifted since the formal review process was enacted, particularly at the program level, as a result of the needs assessment phase of the process, which typically results in one-third fewer programs being included in the full evaluation. See https://eric.ed.gov/?id=EJ1074924.
[16] This total represents the majority of providers in the state. During 2020–21 and 2021–22, DESE will complete the review process for the remaining TPPs.

"not approved" determination for any of the TPPs, 17 TPPs opted to withdraw themselves, effectively ending their approvals, before the conclusion of the process with DESE. In these cases, DESE ceased the evaluation and worked with each individual TPP to establish a timeline for closure that continued to support candidates already in the program.[17]

The approval designations discussed above are important because they are posted publicly and used for accreditation decisions. However, because of the relative lack of variation in approval designations for TPPs *that completed the program review process*, we do not use these approval designations in the remainder of the analysis. Instead, and as described in Section 3, we use the scores associated with each program review domain as our primary measures from the program review process.

3. Data and Measures

All data in this study were provided by DESE. Massachusetts serves approximately one million students and employs roughly 70,000 educators. Massachusetts preparation providers complete more than 5,000 educators annually; about 4,000 of these individuals are initially licensed teacher completers. And about two-thirds of newly minted teachers in Massachusetts are prepared by in-state TPPs. Using information on TPP ratings from 2014–15 to 2019–20, we link these ratings to outcome measures for graduates from these TPPs, as described below.

Program Review Scores

[17] These TPPs are generally small, and when we explored differences in outcomes between graduates of expired TPPs and TPPs that completed the process, we found that we did not have sufficient power to detect even relatively large differences. We therefore focus on TPPs that completed the process for the rest of this analysis. The seven TPPs not considered in this analysis are all going through the review process in the 2020–21 and 2021–22 school years.

The key explanatory variables in this project are the scores received by specific TPPs in the program review process described in Section 2. As noted in that section, only a subset of the 47 TPPs that went through this program review process were evaluated in each year since 2014–15; we refer to the subset of TPPs that underwent program review in a given year as a "review cohort."

As described in Section 2, the DESE review results in TPP ratings on the five organizational-level domains summarized in Table 1. Table 3 shows the distribution of these scores by review cohort and domain. The data that we analyze includes 47 programs that were reviewed: seven programs in the 2014–2015 and 2016–2017 review cohorts; eight programs in the 2015–2016, 2017–2018, and 2019–20 review cohorts; and nine programs in 2018–2019. For the purposes of this study, we convert these ordinal scores to numerical values (Unsatisfactory = 1, Needs Improvement = 2, Proficient = 3, and Exemplary = 4) and add them across domains to calculate a single TPP domain rating. Thus, the minimum score possible is a 5 (if programs receive an "Unsatisfactory" rating on all five organizational domains), and the maximum is 20 (for programs that receive all "Exemplary" ratings). We then standardize these scores across TPPs so the scores have a mean of 0 and a standard deviation of 1. The histogram in Figure 1A illustrates these standardized domain rating scores across TPP graduates.

We also create several alternative measures from the *specific criteria* aligned with each domain. In Table 4, we list the specific criteria associated with the different domains that were consistently used across different phases of program review.[18] Each of the 23 criteria listed in Table 4 are evaluated on a 3-point scale that we also convert to numerical values (1 = Finding, 2 = Met, 3 = Commendation). We create one measure that we call the "average criterion-level

[18] We drop criteria that were added or dropped between phases of program review because we do not observe these criteria for all TPPs.

rating," which is simply the mean score across all of the criteria, standardized across all TPPs; note that this methodology implies equal weighting of criteria that does not necessarily represent how criteria are used in practice. Figure 1B illustrates the distribution of these standardized, criterion-level ratings.

The criterion-level ratings also allow us to explore how the different domains of program review are related to one another. In Table 4, we report the results of a principal components analysis that yielded eight key components with an eigenvalue greater than one. Perfect alignment does not exist between different domains, but the Organization and Continuous Improvement criteria tend to load together (e.g., see components 1 and 2 in Table 4), as do the Partnerships and Field-Based Practices criteria (e.g., see components 3, 4, and 6 in Table 4). The Candidate criteria tend to load on their own principal components (e.g., see component 5 in Table 4). Given that these categories align with prior research on the role of continuous improvement within organizations (e.g., Sessa & London, 2015) and the close connections between partnerships and field-based practices (e.g., St. John, Goldhaber, Krieg, & Theobald, 2018), we therefore create three additional measures that are the average criterion-level ratings for "Organization and Continuous Improvement," "Partnerships and Field-Based Practices," and "Candidate,"—again, standardized across TPPs. The distribution of these measures is shown in Figures 1C–1E.[19]

Outcome Measures

The data discussed above are merged with data on inservice teachers from the state's Education Personnel Information Management System, which include summative performance ratings (SPR). SPR measures are derived from annual teacher evaluation measures; we use either

[19] Correlations between these measures range from 0.55 (between Organization/Continuous Improvement and Candidate) to 0.94 (between average domain-level and average criterion-level ratings).

the raw evaluation scores (measured on a scale of 1–4; henceforth, "SPR ratings") or measures aggregated from the subscores via a graded-response model (henceforth, "SPR GRM ratings"; Kraft, Papay, & Chi, 2020). Using the state's Student Information Management System, the data can be further linked to the demographics and test scores of the students in these teachers' classrooms. We use the SPR measures and student test scores to generate the outcome measures in this study.

Prior research on student test scores (e.g., Koedel et al., 2015) and recent work specifically about SPR in Massachusetts (Cowan, Goldhaber, & Theobald, 2018) has shown considerable variation in these outcomes across different classrooms and schools, so we adjust these measures using standard value-added approaches that regress student achievement or performance ratings Y_{ijt} on student controls X_{ijt}:

$$Y_{ijt} = X_{ijt}\gamma + \epsilon_{ijt} \quad (1)$$

In the model in equation 1, X_{ijt} includes a cubic polynomial in lagged test scores in math and ELA interacted with grade, student demographics, participation in special education or English language learner programs, and classroom and school aggregates of these variables. We additionally include teacher experience, grade-by-grade configuration effects, indicators for membership in a grade involving a structural transition, and indicators for the Partnership for Assessment of Readiness for College and Careers (PARCC) and PARCC online assessments.[20] In models involving teacher evaluations, we additionally include an indicator for a formative assessment and interact grade fixed effects with course subject. We then average residuals from

[20] The structural transition control is an indicator for whether a student's grade is the minimum grade offered in a school. Including this indicator in the models accounts for negative impacts of transitions between school levels on student learning (e.g., Rockoff & Lockwood, 2010).

this regression by teacher and year to construct the measures of teacher effectiveness associated with each outcome.

Samples

The analytic samples for this study are created by merging all of the above data sources with data from the Massachusetts Educator Licensure and Renewal data system. This data system helps teacher licensure applicants create personal profiles and accomplish required tasks via an online portal before obtaining their licenses. We utilize information on program completers from this data system to link teacher candidates to both TPP and the individual information described above, as well as each candidate's year of TPP completion.[21]

Panel A of Table 5 summarizes the number of program completers by year for each review cohort. As the table shows, there are relatively few observations for candidates who completed their program during or after the year of program review, about 25% of the overall sample.[22] Importantly, under DESE's *Guidelines for Program Approval*, program reviews can consider "educator evaluation ratings, program graduates' impact in producing growth in student learning, employment and survey data" for up to three prior cohorts of graduates; outcomes for these cohorts are italicized in Table 5.

Panels B–D of Table 5 provide counts of observations linked to SPR, math value added (VA), and ELA VA measures in each school year and review year. As described above,

[21] Because the graduation months are mainly distributed between April and August for summer completers and winter graduates usually earn their degrees between November of the current year and spring of the next year, the completion year can be more accurately and empirically defined as the 1-year interval between November 1 of this year and October 31 of the next year. The hiring dates for those completers are consistent with this definition of completion year, i.e., graduates usually get hired in a field after they have finished an associated program. Since graduates from one program may attend another program or pursue advanced certification after they have finished one program, we keep each candidate's first completion from a TPP between 2010 and 2019.

[22] This is particularly true for graduates of TPPs in the later review cohorts (these observations are represented in bold), which is not surprising since these graduates would have had less time to participate in the teacher labor market.

outcomes for up to three prior cohorts of graduates (italicized in Table 5) can be used in the program review process, which creates a potential endogeneity issue of candidate outcomes contributing to program review scores (rather than vice versa). We therefore define three samples of completers for this part of the analysis: (a) "All Completers" (i.e., all observations in panels B–D of Table 5); (b) "Not Subject to Review" (i.e., all non-italicized observations in panels B–D of Table 5); and (c) "Post Review" (i.e., all bold observations in panels B–D of Table 5).

Ideally, we would only use the "Post Review" cohorts because the outcomes for these candidates cannot influence their TPP's review scores. However, the issue noted in panel A about the lack of observations from program completers in years during or after the year of program review (shown in bold) is particularly acute when we match these completers with outcomes; for example, about two-thirds of all candidates linked to outcomes in the "Post Review" cohorts come from the first review cohort (2014–15). Therefore, our preferred sample is the "Not Subject to Review" sample, because the sample sizes are relatively large for all review years and this sample removes teachers whose outcomes may have contributed to their provider's program rating.[23]

4. Analytic Approach

The analytic approach for relating program review ratings to the outcome measures described above is as follows. Let O_{ij} be an outcome (SPR or VA) for teacher i who graduated from provider j. In our baseline model, we regress these outcomes against the program review score for provider j, X_j:

[23] We also consider a sub-sample that drops all graduates who graduated more than 5 years before their TPP was reviewed, and results are less precisely estimated but qualitatively similar.

$$O_{ij} = \alpha_0 + \alpha_1 X_j + \varepsilon_{ij}. \qquad (2)$$

The coefficient α_1 can be interpreted as the expected increase in a teacher's outcome associated with a one-point increase in the program review score from that teacher's TPP. In our primary results, we stack the math and ELA VA samples and include a subject indicator in the model in equation 3 to maximize power. The summary statistics in panels A and B of Table 6 provide some *prima facie* evidence of mean differences in these outcomes across quartiles of program review scores.

Identification and inference from the model in equation 2 are complicated by four factors. The first is that teacher candidates non-randomly sort into different TPPs, and as discussed extensively in prior research relating teacher preparation to later teacher outcomes (e.g., Goldhaber et al., 2013; Mihaly et al., 2013), any estimates relating teacher preparation to later teacher outcomes combine both *admission effects* (i.e., differences in the potential effectiveness of candidates who enter different TPPs) and *training effects* (i.e., differences in the quality of training that candidates receive at different TPPs). Most, but not all, of the teachers in the sample took the Massachusetts Tests for Licensure (MTEL) in Communication and Literacy *prior to entering a TPP*.[24] Performance on the MTEL is predictive of teacher effectiveness in Massachusetts (Cowan, Goldhaber, Jin, & Theobald, 2020), and thus scores on these tests allow us to explore some of the implications of differential admissions at TPPs with different program review scores.

Panel C of Table 6 and Figure 2 provide considerable evidence of a relationship between the average MTEL scores of a TPP and the scores that TPPs receive in the program review process. Figure 2, for instance, shows a nearly linear, positive relationship between a TPP's

[24] Among teachers with non-missing program entry dates, 90% of these teachers took the MTEL Communication and Literacy tests prior to program entry.

average MTEL scores and the TPP's review score. This suggests that any relationship between program review scores and later teacher outcomes like SPR or VA may be driven, in part, by admission effects; i.e., candidates with lower MTEL scores are more likely to attend TPPs with lower program review scores and also tend to be less effective teachers. This is not necessarily a problem for statistical inference from the model in equation 2; i.e., from the state's perspective, it may not matter whether a relationship between program review scores and later outcomes is driven by the candidates that a TPP selects or the training that candidates receive at TPPs (i.e., the developmental effects of attending a TPP), because either might result in a finding of more effective teachers graduating from higher rated TPPs. But in other contexts it could matter—for example, if there is a desire to better understand the training at TPPs specifically. With this in mind, we estimate some specifications of the model in equation 2 that control for a teacher's MTEL scores, $MTEL_i$:

$$O_{ij} = \beta_0 + \beta_1 X_j + \beta_2 MTEL_i + \varepsilon_{ij}. \quad (3)$$

The coefficient β_1 in equation 3 can be interpreted as the relationship between program review scores and a teacher's outcome accounting for baseline differences in teachers' MTEL scores. More directly, differences between the estimated coefficients $\hat{\alpha}_1$ from equation 2 and $\hat{\beta}_1$ from equation 3 provide some evidence about the extent to which relationships between program review scores and SPR are sensitive to the sorting of candidates to TPPs as captured by MTEL scores. However, it is worth noting that we cannot adjust for selection into training programs based on unobservable factors. Adding the MTEL scores to regressions predicting teacher effectiveness measures increases the R^2 by only negligible amounts, which suggests that there is still considerable scope for non-random sorting into TPPs along unobservable dimensions to influence these results.

The second issue with inference from the models in equations 2 and 3 is that teacher candidates non-randomly enter the state teaching workforce in Massachusetts. This is only a problem if (a) candidates from TPPs with different program review scores are differentially likely to be observed in the workforce, and (b) there are unobserved factors that predict both workforce entry and teacher outcomes (i.e., sample selection bias). We are unable to test the second possibility, but we do find some evidence of differential selection into the sample by program review scores (results are available from the authors upon request). Specifically, completers from mid-scoring TPPs tend to be employed less frequently in Massachusetts public schools than completers from high-scoring and low-scoring TPPs. This could lead to sample selection bias if candidates with more teaching potential are more likely to obtain employment in the state's public K–12 schools.[25] We are limited in our ability to account for this potential source of bias, so simply note this as one reason we interpret our estimates in purely descriptive terms.

The third complication is that teachers who enter the workforce are non-randomly distributed across different classrooms and schools in the state (see Cowan et al., 2017, for evidence of this in Massachusetts). If a given TPP disproportionately sends teachers to a particularly effective or ineffective school or district, for example, naïve models will attribute these school effects to the TPPs (and thus to the scores they receive in program review). Our primary approach to this potential source of bias is to consider SPR and VA measures that are adjusted with a school or district fixed effect; these measures make comparisons only between teachers within the same school or district and thus account for this potential source of

[25] Graduates from some TPPs may differentially choose to teach in another state, which can be observed in employment rates for various TPPs available on DESE's public profiles, though that alone may not explain the effect we observe here.

confounding. However, as discussed in Boyd and colleagues (2009), Goldhaber and associates (2013), and Goldhaber and Ronfeldt (2020), this approach may also remove some of the true differences in teacher effectiveness across different schools and district. Suppose, for example, that a given school hires only the best graduates of ineffective TPPs and the lower performing graduates of effective TPPs. When the outcome includes school fixed effects, the differences between these TPPs will look smaller than they are in real life. Since it is difficult to disentangle these two competing issues (non-random sorting across schools and true differences in teacher quality across schools) with short panels of observational data, we report estimates from models in which the outcomes do and do not account for school or district fixed effects, and interpret the results accordingly.

The final issue with statistical inference in this context is quantifying uncertainty in the estimates from the models in equations 2 and 3. Specifically, given that all graduates from the same provider *j* have the same value of X_j in these models, naïve standard errors may overstate our confidence in these estimates. The typical approach in this situation is to cluster standard errors at the program level, but as shown in Cameron and colleagues (2008), analytic approaches to clustered standard errors perform poorly when the number of clusters is low, as in this application. We therefore follow the recommendation of Cameron and associates (2008) and conduct inference using the wild cluster bootstrap using the *boottest* command in STATA (Roodman, Nielsen, MacKinnon, & Webb, 2019); we present both the bootstrapped standard error and associated 95% confidence interval for each estimate, which account for additional uncertainly associated with the clustering of observations within a small number of clusters (i.e., TPPs).

5. Results

Table 7 summarizes the estimates from the models in equations 3 and 4 in which the outcomes are regression-adjusted SPR.[26] The estimates in column 1 are from the base model in equation 3, column 2 presents estimates from the model in equation 4 that controls for candidates' licensure test scores on the MTEL Communication and Literacy fields, columns 3 and 4 present estimates from analogous models in which the SPR measures are adjusted by district fixed effects as described in Section 4, and the SPR measures in columns 5 and 6 are adjusted by school fixed effects. The first estimate in column 1 shows that a one standard deviation increase in a provider's review score is predictive of a 0.0547 standard deviation increase in the adjusted SPR of the average graduate from that provider; this positive, largely linear relationship can be seen in the first scatterplot in **Figure 3**. To put the magnitude of this relationship in context, this is roughly 20% of the average increase in SPR between a first- and second-year teacher.

The relationships between program review scores and SPR are marginally significant when the program review scores are aggregated across domains (panel 1) and significant at the 0.05 level when scores are aggregated across criteria (panel 2) and for two subcategories described in Section 4 ("Partnerships and Field Based Practices" and "Candidate"). These relationships attenuate somewhat but remain statistically significant when we control for teachers' MTEL scores in column 2. This could be due to a number of factors; e.g., Massachusetts' program review process may capture, in part, TPP admissions as captured by incoming MTEL scores (as shown in Figure 2).

[26] For parsimony, we present estimates from SPR aggregated via a graded-response model (GRM) as described in Section, but results are qualitatively similar for the raw SPR measures.

The relationships between program review scores and SPR attenuate even more when the SPR measures are adjusted for school and district effects (columns 3–6) and are not consistently significant at conventional levels, with the exception of the "Partnerships and Field-Based Practices" subcategory. The SPR measures adjusted by school effects are our preferred outcome measure because prior work in Massachusetts (Cowan et al., 2018) found that this specification results in unbiased, out-of-sample predictions of teachers' SPR. Thus, our overall conclusion is that program review scores are modest and sometimes statistically significant predictors of TPP graduates' SPR ratings.

Table 8 presents similar specifications in which the outcomes are TPP graduates' value added (stacked across math and ELA). The overall patterns are similar for value added for SPR, as program review scores are positive and consistently statistically significant predictors of TPP graduates' value added in both subjects, even controlling for differences in MTEL scores across TPPs. In both subjects, a one standard deviation increase in a provider's review score is predictive of a 0.02–0.03 standard deviation increase in student test-score gains in the classroom of the average graduate from that provider. This difference in value added is about half of the expected returns to the first year of teaching experience in the state.

That said, the relationships between program review scores and value added attenuates considerably when the value-added measures are adjusted for district and school effects. This suggests that the relationships between program review scores and TPP graduates' value added are driven primarily by the sorting of graduates from TPPs with higher program review scores to schools and districts with greater learning gains.[27] The results for the average domain-level

[27] We do find some evidence of non-random sorting of graduates of higher scoring TPPs to more advantaged classrooms; for example, students eligible for free or reduced-priced lunch are more likely to have teachers from low-scoring TPPs. This sorting along observed dimensions is not an issue for inference, though, since we control directly for these relationships in the models in equations 3 and 4.

rating, for example, suggest that about two-thirds of the relationship between program review scores and value added is explained by sorting across districts, while over 80% is explained by sorting across schools.

However, unlike with SPR, where there is a clear preferred specification, it is not clear which specification is preferable for assessing the relationships between program review scores and TPP graduates' contributions to student learning gains. The fundamental issue is discussed in Section 4—namely, that it is difficult to disentangle non-random sorting across schools and districts and true differences in teacher quality across schools with relatively short panels of observational data. Thus, the interpretation of this result comes down to a relatively simple question that cannot be answered with available data: If a school has strong test-score gains and a large number of graduates from TPPs with high program review scores, is this due to school-level factors outside of these teachers' control, or does the school have strong test-score gains precisely because it employs a large number of graduates from high-scoring TPPs? The former explanation would suggest that the fixed-effects results suggesting little relationship between program reviews scores and value added are preferable, as they do not misattribute school-level factors to TPP review scores. On the other hand, the latter explanation would suggest that the models without fixed effects suggesting significant relationships between review scores and value added are preferable, because they appropriately attribute differences across different schools and districts to the composition of their teaching staffs. It is worth stressing again that we cannot test these competing explanations with available data, so as we discuss in the conclusion, we urge further research that may be able to disentangle TPP and school and district contributions to student learning.

6. Conclusions

This paper represents the first attempt to connect ratings generated through a TPP review process to educator outcomes. As such, we view this as a preliminary investigation that directly informs future research in this area. Moreover, given the distinct approach to program review implemented by DESE and the significant relationships between program review scores and teacher effectiveness in some specifications, we believe that there are important implications in Massachusetts and nationally. First, given the investment of resources—on the part of both the state and the TPPs—it is important to recognize that the judgments made about TPPs through the review process do provide some signal of educator effectiveness to potential candidates or hiring districts, particularly in the Partnerships and Field-Based Practices domain. As a result, DESE may want to consider drawing additional attention to these areas of program review. Nationally, the specific approach to reviews in Massachusetts may be worthy of consideration for other states, and may warrant additional study. At the very least, we would encourage similar studies to be completed on program review efforts in other states or by national accrediting agencies, given that the review process in other states often looks quite different from the review process considered in this paper.

In particular, it is worth noting that a feature of the Massachusetts process is its emphasis at the organization level, not the discrete program level. Findings from this analysis suggest that this focus may be sufficient and appropriate for state-level decision making, as opposed to the more time-intensive and complex endeavor of programmatic-level reviews. Alternatively, the scores from this process may have had more predictive power if they had been generated in a more targeted way at the program level; in particular, additional interrogation at the program level may provide insight into the strong connection between candidates' performance on the

MTEL and judgments made through the review. Evidence from other states with different program review processes will shed light on these different hypotheses.

The non-random sorting of candidates from high-performing TPPs into high-performing schools is an important caveat to these findings to be explored further. There are several explanations for this sorting that may be independent of TPPs (e.g., school-based hiring practices, induction and mentoring programs), or it may be that TPPs are judged as high-performing precisely *because* they are in close partnership with these high-performing schools. This last hypothesis is consistent with a policy push from DESE over the last several years (Massachusetts Department of Elementary and Secondary Education, 2018) and is a focus of the current structure of reviews. In fact, TPPs that achieved the state's highest rating of "Approved with Distinction" did so with highest marks ("Exemplary") in the Partnerships and Field-Based Experiences domain, signaling exactly the type of intentional pipeline and relationship to the PK–12 schools and districts explored through this inquiry. Evidence from other states, potentially with program review processes spanning more years or less clear sorting from specific TPPs into specific schools and districts, will be important to better understand TPP and school/district contributions to student learning.

A final big question not answered through this study is whether the quality of preparation *improved* in Massachusetts as a result of or in relation to the implementation of the DESE review process. In particular, Massachusetts set out to narrow the variation across programs in a TPP and ensure some consistency within a provider through its review effort; this additional analysis would be particularly informative given the debate about whether to focus at the program or organizational level in these reviews. Future work could also leverage data on organizations that

have made organizational and programmatic changes based on review findings to assess changes in TPP practices in response to program review.

Finally, it is significant that the field of providers has been reduced by almost a quarter (17 out of the original 71 TPPs closed) without the significant political upheaval typically experienced if/when a TPP is not approved through reviews. Yet, given the small size of many of these TPPs, we are not able to determine the outcomes of completers from those programs to provide a solid comparison to those that continued through the full process. This is another area of investigation that is still open and a potential avenue for future research.[28]

[28] See https://www.doe.mass.edu/edprep/domains/improvement/improvement.html.

References

Bastian, K. C., Lys, D., & Pan, Y. (2018). A framework for improvement: Analyzing performance-assessment scores for evidence-based teacher preparation program reforms. *Journal of Teacher Education*, *69*(5), 448–462. https://doi.org/10.1177/0022487118755700

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440. https://doi.org/10.3102/0162373709353129

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414–427.

CCSSO. (2012). *Our responsibility, our promise: Transforming educator preparation and entry into the profession*. Council of Chief State School Officers (CCSSO). https://ccsso.org/sites/default/files/2017-10/Our%20Responsibility%20Our%20Promise_2012.pdf

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. https://doi.org/10.1257/aer.104.9.2593

Conaway, C., & Goldhaber, D. (2020). Appropriate standards of evidence for education policy decision making. *Education Finance and Policy*, *15*(2), 383–396. https://doi.org/10.1162/edfp_a_00301

Cowan, J., Goldhaber, D., & Theobald, R. (2017). *Massachusetts educator preparation and licensure: Year 1 report*. American Institutes for Research. https://www.doe.mass.edu/research/reports/2017/05EdPrep-Year1Report.pdf

Cowan, J., Goldhaber, D., & Theobald, R. (2018). *Performance evaluations as a measure of teacher effectiveness when standards differ: Accounting for variation across classrooms, schools, and districts*. CALDER Working Paper No. 197-0618-2. https://caldercenter.org/sites/default/files/CALDER%20WP%20197-0618-2.pdf

Cowan, J., Goldhaber, D., Jin, Z., & Theobald, R. (2020). *Teacher licensure tests: Barrier or predictive tool?* CALDER Working Paper No. 245-1020. https://caldercenter.org/sites/default/files/WP%20245-1020_0.pdf

Massachusetts Department of Elementary and Secondary Education. (2018). Educator preparation formal review advisory: Partnerships domain. https://www.doe.mass.edu/edprep/resources/guidelines-advisories/partnerships-advisory.pdf

Massachusetts Department of Elementary and Secondary Education. (2020). Educator preparation review and approval—Educator preparation. https://www.doe.mass.edu/edprep/review/

Feuer, M. J., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options*. National Academy of Education. https://naeducation.org/wp-content/uploads/2016/11/028489-Evaluation-of-Teacher-prep.pdf

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources*, *XLII*(4), 765–794. https://doi.org/10.3368/jhr.XLII.4.765

Goldhaber, D. (2019). Evidence-based teacher preparation: Policy context and what we know. *Journal of Teacher Education*, *70*(2), 90–101. https://doi.org/10.1177/0022487118800712

Goldhaber, D., & Koedel, C. (2019). Public accountability and nudges: The effect of an information intervention on the responsiveness of teacher education programs to external ratings. *American Educational Research Journal*, *56*(5), 1557–1589. https://doi.org/10.3102/0002831218820863

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29–44. https://doi.org/10.1016/j.econedurev.2013.01.011

Goldhaber, D., & Ronfeldt, M. (2020). Toward causal evidence on effective teacher preparation. In *Linking teacher preparation program design and implementation to outcomes for teachers and students* (pp. 211–236). Information Age Publishing.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199–208. https://doi.org/10.1080/09645299900000018

Greenberg, J., Pomerance, L., & Walsh, K. (2011). *Student teaching in the United States*. National Council on Teacher Quality. https://eric.ed.gov/?id=ED521916

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271. https://doi.org/10.1257/aer.100.2.267

Holmes Group, Inc. (1986). *Tomorrow's teachers: A report of the Holmes Group*. East Lansing, MI: Author. https://eric.ed.gov/?id=ED270454

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107. https://doi.org/10.1086/699018

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs? *Education Finance and Policy*, *10*(4), 508–534. https://doi.org/10.1162/EDFP_a_00172

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1–36. https://doi.org/10.3368/jhr.54.1.0916.8265R3

Kraft, M. A., Papay, J. P., & Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315–347. https://doi.org/https://doi.org/10.1002/pam.22193

Levine, A. (2006). *Educating school teachers*. The Education Schools Project. http://edschools.org/pdf/Educating_Teachers_Report.pdf

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. https://doi.org/10.1162/EDFP_a_00110

National Academies of Sciences, Engineering, and Medicine. (2020). *Changing expectations for the K-12 teacher workforce: Policies, preservice education, professional development, and the workplace*.

National Commission on Teaching & America's Future. (1996). *What matters most: Teaching for America's Future*. Report of the National Commission on Teaching & America's Future. New York, NY: Author. https://eric.ed.gov/?id=ED395931

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237

Rockoff, J., & Lockwood, B. (2010). Stuck in the middle: Impacts of grade configuration in public schools. *Journal of Public Economics*, *94*(11–12), 1051–1061.

Ronfeldt, M., & Campbell, S. L. (2016). Evaluating teacher preparation using graduates' observational ratings. *Educational Evaluation and Policy Analysis*, *38*(4), 603–625. https://doi.org/10.3102/0162373716649690

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal*, *19*(1), 4–60. https://doi.org/10.1177/1536867X19830877

Sessa, V. I., & London, M. (2015). *Continuous learning in organizations: Individual, group, and organizational perspectives*. Psychology Press.

St. John, E., Goldhaber, D., Krieg, J., & Theobald, R. (2018). *How the match gets made: Exploring student teacher placements across teacher education programs, districts, and schools*. CEDR Working Paper No. 10052018-1-1. National Center for Analysis of Longitudinal Data in Education Research (CALDER).

U.S. Department of Education. (2014). *Federal register, volume 79 issue 122*. https://www.govinfo.gov/content/pkg/FR-2014-06-25/html/2014-14783.htm

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*, *64*, 298–312. https://doi.org/10.1016/j.econedurev.2018.01.005

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, *53*, 31–45. https://doi.org/10.1016/j.econedurev.2016.05.002

Zeichner, K. (2006). Reflections of a university-based teacher educator on the future of college and university-based teacher education. *Journal of Teacher Education, 57*(3), 326–340.

Table 1. Program review domains and guiding questions

Domain	Guiding Question
The Organization	Is the organization set up to support and sustain effective preparation programs?
Partnerships	Is educator preparation from your organization meeting the needs of the PK–12 system?
Continuous Improvement	Is your organization driven by continuous improvement efforts that result in better prepared educators?
Candidate	Is the candidate's experience throughout the program contributing to effective preparation?
Field-Based Experiences	Do candidates have the necessary experiences in the field to be ready for the licensure role?
Instruction	Do candidates have the necessary knowledge and skills to be effective?

Table 2. Approval designations in program review data

Approval Designation following Formal Review	**Number of TPPs**
Approved with Distinction	4
Approved	31
Approved with Conditions	8
Probationary Approval	4
Not Approved	0
Exited the process	17

Table 3. Review scores by review year and domain

Panel A: The Organization

	Review Year						
	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total
Exemplary	1	1	1	0	0	2	5
Proficient	3	5	4	8	6	5	31
Needs Improvement	0	1	2	0	2	0	5
Unsatisfactory	3	1	0	0	1	1	6
Total	7	8	7	8	9	8	47

Panel B: Partnerships

	Review Year						
	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total
Exemplary	1	1	2	0	0	1	5
Proficient	3	5	3	4	5	7	27
Needs Improvement	2	2	2	4	3	0	13
Unsatisfactory	1	0	0	0	1	0	2
Total	7	8	7	8	9	8	47

Panel C: Continuous Improvement

	Review Year						
	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total
Exemplary	1	0	1	0	0	2	4
Proficient	1	5	5	7	4	5	27
Needs Improvement	2	2	1	1	5	1	12
Unsatisfactory	3	1	0	0	0	0	4
Total	7	8	7	8	9	8	47

Panel D: Candidate

	Review Year						
	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total
Exemplary	0	0	1	1	0	1	3
Proficient	4	8	5	6	4	5	32
Needs Improvement	2	0	1	1	2	2	8
Unsatisfactory	1	0	0	0	3	0	4
Total	7	8	7	8	9	8	47

Panel E: Field-Based Experiences

	Review Year						
	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total
Exemplary	0	0	1	0	2	1	4
Proficient	3	5	3	6	2	5	24
Needs Improvement	3	2	3	2	4	2	16
Unsatisfactory	1	1	0	0	1	0	3
Total	7	8	7	8	9	8	47

Table 4. Principal components analysis of criterion scores

Domain	Criteria	C1	C2	C3	C4	C5	C6	C7	C8
The Organization	Organizational structure demonstrates sufficient capacity to carry out responsibility and decision making for educator preparation programs.	0.215	0.180	0.093	0.130	0.033	-0.262	-0.023	-0.021
	Systems/structures that support collaboration within departments and across disciplines improve candidate preparation.	**0.489**	-0.143	0.062	-0.091	0.129	0.007	0.164	-0.109
	Budget supports ongoing sustainability and allocates resources according to organizational goals.	0.121	0.243	0.194	-0.154	-0.011	-0.042	0.178	0.041
	All candidates, regardless of program or delivery model, have equitable and consistent access to resources.	-0.004	0.156	-0.182	**0.567**	-0.011	-0.136	0.091	0.021
	Recruitment, selection, and evaluation processes result in the hiring and retention of effective faculty and staff.	**0.374**	0.019	-0.092	0.195	0.221	-0.197	0.023	0.068
	Faculty/instructors and staff engage in professional development and work in the field that has clear application to preparing effective educators.	-0.007	**0.327**	0.191	-0.136	0.196	-0.018	0.102	-0.251
Partnerships	Partners make contributions that inform Sponsoring Organization's continuous improvement efforts.	0.038	-0.002	**0.461**	0.020	**-0.395**	-0.077	-0.028	-0.013
	Partnerships improve experience for preparation candidates and outcomes for PK–12 students.	0.043	0.012	0.081	**0.489**	-0.037	0.239	-0.123	-0.092
	Sponsoring Organization responds to district/school needs through focused recruitment, enrollment, retention, and employment (e.g., placement agreement with local district) efforts.	-0.043	-0.008	0.435	0.199	-0.140	-0.102	-0.045	0.245
	Sponsoring Organizations evaluate partnerships on an ongoing basis, sustain those that are effective, and take steps to improve those that are not.	0.162	0.225	0.050	-0.145	-0.198	**0.317**	-0.190	0.024
Continuous Improvement	The consistent and ongoing use of internal and external evidence, including elementary and secondary education data, informs strategic decisions that impact the Sponsoring Organization, the education programs, candidates, and employing organizations.	-0.196	**0.637**	-0.129	0.097	0.027	-0.063	0.004	-0.077
	Sponsoring Organization acts on feedback solicited from internal and external stakeholders (including candidates, graduates, district and school personnel and employers) in continuous improvement efforts.	**0.359**	0.097	-0.104	0.101	0.016	0.146	-0.117	-0.029
Candidate	Recruitment efforts yield a diverse candidate pool.	-0.185	0.144	0.039	-0.017	0.286	0.241	0.153	**0.317**
	Admission criteria and processes are rigorous such that those admitted demonstrate success in the program and during employment in a licensure role.	0.070	0.016	0.071	0.105	**0.446**	0.141	-0.027	0.228
	Candidates at risk of not meeting standards are identified throughout the program (in pre-practicum, during coursework, and while in practicum) and receive necessary supports and guidance to improve or exit.	0.240	0.052	-0.022	-0.007	**0.597**	-0.019	-0.146	0.006
	Waiver policy ensures that academic and professional standards of the licensure role are met.	0.004	-0.031	-0.022	0.005	-0.092	0.078	**0.776**	-0.041
Field-Based Experiences	Practicum hours meet regulatory requirements as per 603 CMR 7.04 (4).	0.021	-0.089	0.005	-0.094	0.072	-0.158	-0.024	**0.812**
	District partners are involved in the design, implementation, and assessment of field-based experiences.	-0.001	-0.102	**0.609**	-0.078	0.064	-0.072	-0.018	-0.031
	Responsibilities in field-based experiences build to candidate readiness for full responsibility in licensure role.	**0.585**	-0.270	0.010	-0.061	0.084	0.181	-0.015	0.082
	Candidates participate in field-based experiences that cover the full academic year.	0.101	-0.016	-0.109	0.047	0.044	**0.745**	0.106	-0.153
	Field-based experiences are embedded in program coursework.	0.133	0.031	-0.037	0.127	-0.326	0.132	**0.483**	0.100
	Supervising Practitioners and Program Supervisors receive training, support, and development from the Sponsoring Organization that impacts candidate effectiveness.	-0.125	**0.508**	-0.026	0.026	0.056	0.118	-0.091	0.006
	Field-based experiences are in settings with diverse learners (e.g., students from diverse ethnic, racial, gender, socioeconomic, and exceptional groups).	-0.108	-0.135	0.279	**0.473**	0.148	0.108	0.047	-0.224

Note. Table shows all factors with an eigenvalue greater than 1. Factor loadings with an absolute value greater than 0.3 are in bold.

Table 5: Number of program completers by sample and review cohort

Panel A: All Completers	**Review Cohort**						
	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total
2010	644	251	586	214	245	403	2,343
2011	607	281	616	283	205	490	2,482
2012	*644*	228	640	269	211	554	2,546
2013	*677*	*192*	638	217	292	466	2,482
2014	*670*	*220*	*535*	288	281	517	2,511
2015	**664**	*181*	*472*	*230*	299	435	2,281
2016	**648**	**204**	*511*	*227*	*265*	468	2,323
2017	**636**	**167**	**456**	*227*	*279*	*460*	2,225
2018	**590**	**174**	**414**	**180**	*252*	*428*	2,038
2019	**580**	**199**	**402**	**159**	**277**	*475*	2,092
Total	6,360	2,097	5,270	2,294	2,606	4,696	23,323
Panel B: Summative Performance Rating Sample							
2010	217	104	254	51	99	138	863
2011	251	124	314	92	88	169	1,038
2012	*273*	129	285	80	81	206	1,054
2013	*301*	*81*	306	71	123	181	1,063
2014	*346*	*120*	*244*	91	107	231	1,139
2015	**310**	*89*	*205*	*71*	104	191	970
2016	**284**	**105**	*240*	*65*	*98*	167	959
2017	**266**	**70**	**172**	*57*	*73*	*182*	820
2018	**228**	**58**	**109**	**43**	*70*	*116*	624
2019	**128**	**29**	**31**	**12**	**33**	*79*	312
Total	2,604	909	2,160	633	876	1,660	8,842
Panel C: Math Value-Added Sample							
2010	52	26	64	22	35	28	227
2011	83	26	73	35	24	45	286
2012	*70*	34	73	19	24	60	280
2013	*63*	*12*	81	18	39	44	257
2014	*80*	*36*	*73*	27	29	70	315
2015	**54**	*16*	*44*	*28*	30	41	213
2016	**62**	**18**	*47*	*16*	*24*	41	208
2017	**55**	**16**	**42**	*13*	*22*	*39*	187
2018	**62**	**10**	**19**	**8**	*18*	*34*	151
2019	**38**	**3**	**9**	**3**	**13**	*20*	86
Total	619	197	525	189	258	422	2,210
Panel D: English Language Arts Value-Added Sample							
2010	52	25	65	23	28	33	226
2011	81	29	80	34	23	41	288
2012	*63*	26	83	21	26	53	272
2013	*76*	*21*	82	23	32	37	271
2014	*95*	*31*	*76*	17	25	58	302
2015	**60**	*16*	*52*	*27*	24	35	214
2016	**66**	**15**	*52*	*16*	*22*	29	200
2017	**58**	**17**	**47**	*11*	*14*	*38*	185
2018	**45**	**11**	**26**	**10**	*20*	*25*	137
2019	**35**	**7**	**7**	**2**	**7**	*22*	80
Total	631	198	570	184	221	371	2,175

Note. Bold cells identify post-review cohorts; all cells not italicized indicate non-subject-to-review cohorts.

Table 6. Descriptive statistics by quantile of program review score

Panel A. Average SPR by Overall Program Rating Quantiles

Quantile	Overall Rating	Overall Rating w/ District Effect	Overall Rating w/ School Effect	Overall GRM Rating	Overall GRM Rating w/ District Effect	Overall GRM Rating w/ School Effect
0%-25%	-0.0276	-0.0175	-0.0171	-0.0668	-0.0417	-0.0432
25%-50%	-0.0091	-0.0112	-0.0143	-0.0326	-0.0383	-0.0423
50%-75%	-0.0089	-0.0048	-0.0110	-0.0078	-0.0001	-0.0144
75%-100%	0.0332	0.0152	0.0159	0.1011	0.0554	0.0513

Panel B. Average VA by Overall Program Rating Quantiles

Quantile	Math VA	Math VA w/ District Effect	Math VA w/ School Effect	ELA VA	ELA VA w/ District Effect	ELA VA w/ School Effect
0%-25%	-0.0479	-0.0175	-0.0160	-0.0503	-0.0168	-0.0159
25%-50%	-0.0008	-0.0131	-0.0134	0.0339	-0.0027	-0.0008
50%-75%	-0.0118	-0.0061	-0.0125	0.0011	-0.0171	-0.0203
75%-100%	0.0421	-0.0054	-0.0074	0.0298	-0.0113	-0.0088

Panel C. Average MTEL Score by Overall Program Rating Quantiles

Quantile	Standardized MTEL Reading Score	Standardized MTEL Writing Score
0%-25%	0.0371	0.0193
25%-50%	0.2333	0.3180
50%-75%	0.1498	0.2786
75%-100%	0.4476	0.5911

Notes. ELA = English language arts; GRM = graded-response model. MTEL = Massachusetts Tests for Educator Licensure; SPR = summative performance rating; VA = value added.

Table 7. Regressions predicting SPR as function of overall program review scores (Not subject to review sample)

	SPR GRM Rating	SPR GRM Rating	SPR GRM Rating w/ District Effect	SPR GRM Rating w/ District Effect	SPR GRM Rating w/ School Effect	SPR GRM Rating w/ School Effect
Average Domain-Level Rating	.0547+	.0443+	.0359	.0259	.0325	.0225
95% Confidence Interval	(-0.0158, 0.1736)	(-0.0108, 0.1559)	(-0.0213, 0.1319)	(-0.0393, 0.1160)	(-0.0256, 0.1243)	(-0.0165, 0.1125)
Bootstrapped *P*-value	.0741	.0611	.1041	.1652	.1311	.1411
Average Criterion-Level Rating	.0541+	.0454*	.0334+	.025	.0302+	.0214
95% Confidence Interval	(-0.0042, 0.1403)	(0.0089, 0.1258)	(-0.0184, 0.1009)	(-0.0207, 0.0886)	(-0.0105, 0.1020)	(-0.0098, 0.0819)
Bootstrapped *P*-value	.0531	.034	.0791	.1321	.0911	.1532
Organization and Continuous Improvement	.0425+	.034+	.0245	.0164	.0219	.0136
95% Confidence Interval	(-0.0004, 0.1321)	(-0.0032, 0.1155)	(-0.0329, 0.0969)	(-0.0338, 0.0794)	(-0.0262, 0.0901)	(-0.0320, 0.0713)
Bootstrapped *P*-value	.0531	.0611	.1922	.3233	.1642	.3483
Partnerships and Field-Based Practices	.062*	.0533*	.0412*	.0329+	.0371*	.0284+
95% Confidence Interval	(0.0086, 0.1406)	(0.0147, 0.1299)	(0.0056, 0.1055)	(-0.0009, 0.0932)	(0.0033, 0.1013)	(-0.0011, 0.0902)
Bootstrapped *P*-value	.044	.022	.042	.0541	.042	.0511
Candidate	.069*	.057*	.0403+	.028	.0369*	.024
95% Confidence Interval	(0.0213, 0.1410)	(0.0103, 0.1288)	(-0.0033, 0.0934)	(-0.0159, 0.0783)	(0.0014, 0.0919)	(-0.0079, 0.0743)
Bootstrapped *P*-value	.015	.029	.0571	.1231	.044	.1461
MTEL Controls	N	Y	N	Y	N	Y
Observations	6,132	6,106	6,132	6,106	6,132	6,106

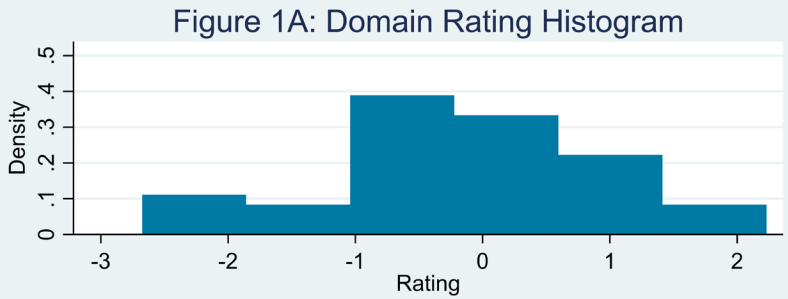
Notes. GRM = graded-response model; MTEL = Massachusetts Tests for Educator Licensure; SPR = summative performance rating. *P*-values from cluster wild bootstrap (clustered by provider): +$p < 0.10$; *$p < 0.05$.

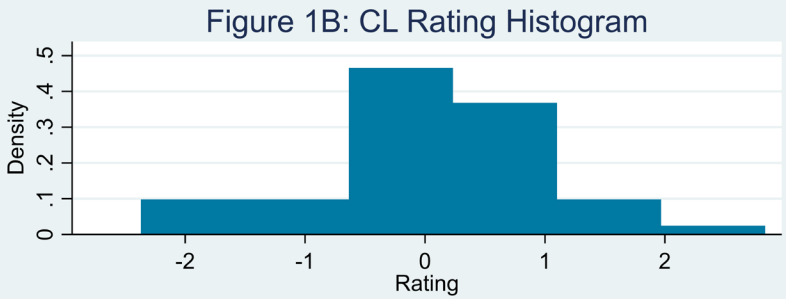
Table 8. Regressions predicting math and English language arts stacked value added as function of overall program review scores (Not subject to review sample)

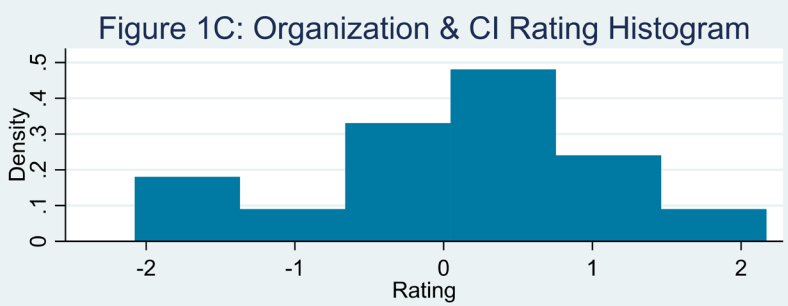
	Stacked VA	Stacked VA	Stacked VA w/ District Effect	Stacked VA w/ District Effect	Stacked VA w/ School Effect	Stacked VA w/ School Effect
Average Domain-Level Rating	.0262+	.0235+	.0075	.0071	.0038	.0036
95% Confidence Interval	(-0.0002, 0.0660)	(-0.0017, 0.0603)	(-0.0057, 0.0130)	(-0.0061, 0.0123)	(-0.0101, 0.0123)	(-0.0098, 0.0112)
Bootstrapped *P*-value	.0501	.0621	.1742	.1992	.3794	.4234
Average Criterion-Level Rating	.0233*	.021*	.005	.0046	.0022	.0021
95% Confidence Interval	(0.0042, 0.0535)	(0.0072, 0.0484)	(-0.0066, 0.0089)	(-0.0084, 0.0085)	(-0.0091, 0.0073)	(-0.0096, 0.0069)
Bootstrapped *P*-value	.031	.018	.2713	.3744	.4765	.4945
Organization and Continuous Improvement	.0239*	.0218*	.0057	.0054	.0027	.0027
95% Confidence Interval	(0.0053, 0.0558)	(0.0041, 0.0543)	(-0.0059, 0.0106)	(-0.0082, 0.0102)	(-0.0081, 0.0080)	(-0.0087, 0.0077)
Bootstrapped *P*-value	.018	.029	.2112	.2482	.4194	.3904
Partnerships and Field-Based Practices	.0238*	.0215*	.0063	.006	.0035	.0034
95% Confidence Interval	(0.0047, 0.0498)	(0.0006, 0.0518)	(-0.0058, 0.0103)	(-0.0055, 0.0095)	(-0.0072, 0.0086)	(-0.0070, 0.0083)
Bootstrapped *P*-value	.034	.044	.1502	.1592	.2803	.3233
Candidate	.0269*	.0231*	.002	.0012	-.0009	-.0013
95% Confidence Interval	(0.0069, 0.0500)	(0.0014, 0.0457)	(-0.0150, 0.0105)	(-0.0176, 0.0107)	(-0.0177, 0.0075)	(-0.0189, 0.0072)
Bootstrapped *P*-value	.019	.042	.7558	.8709	.8529	.7838
MTEL Controls	N	Y	N	Y	N	Y
Observations	3,268	3,252	3,268	3,252	3,268	3,252

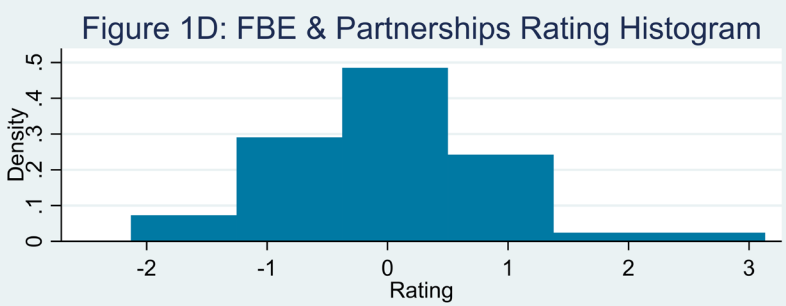
Notes. MTEL = Massachusetts Tests for Educator Licensure; VA = value added. *P*-values from cluster wild bootstrap (clustered by provider): +$p < 0.10$; *$p < 0.05$.

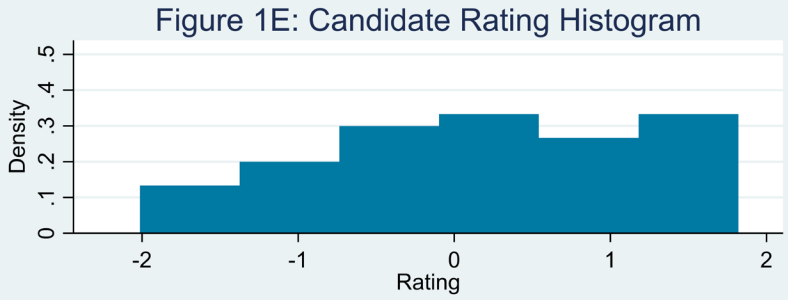
Figure 1. Distribution of program review scores, by institution





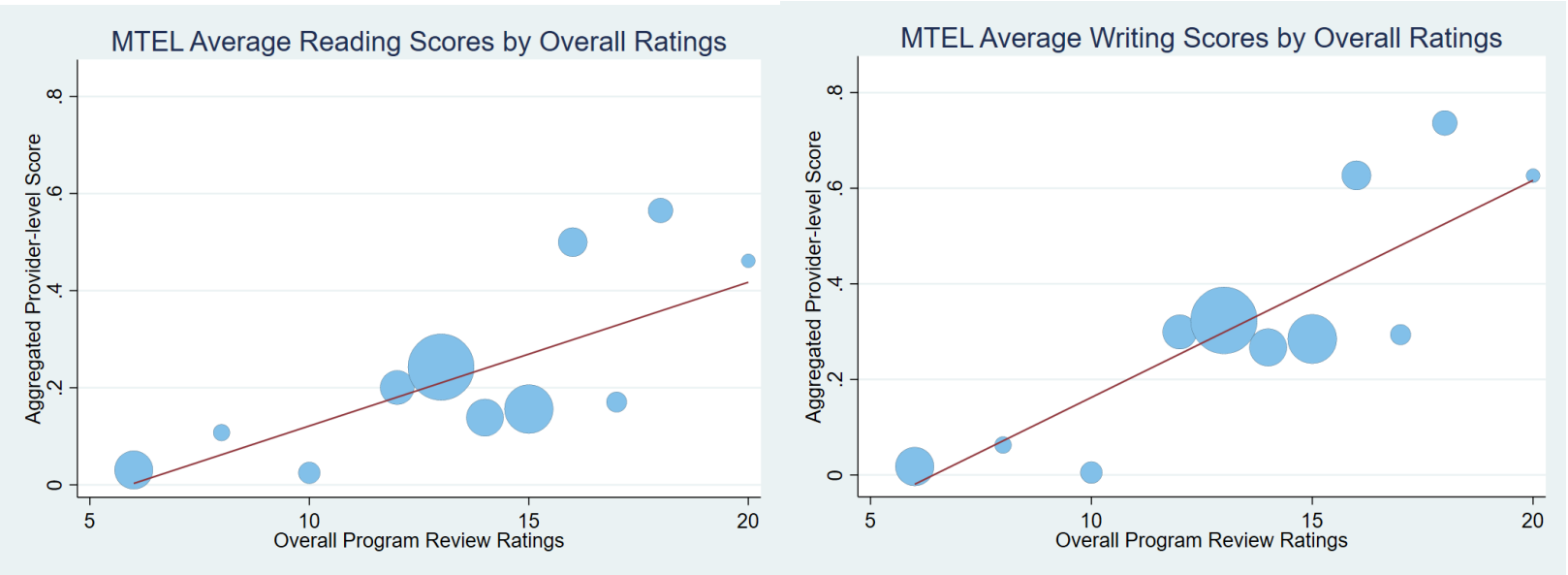






Note. CL = criterion level; CI = confidence interval; FBE = field based experiences.

Figure 2. Average MTEL Communication and Literacy test scores, by overall ratings



Note. MTEL = Massachusetts Tests for Educator Licensure.

Figure 3. Provider-level outcomes, by program review score