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

From the Clinical Experience to the Classroom: Assessing the Predictive Validity of the Massachusetts Candidate Assessment of Performance

Bingjie Chen

James Cowan

Dan Goldhaber

Roddy Theobald

WORKING PAPER No. 223-1019 • October 2019

From the Clinical Experience to the Classroom: Assessing the Predictive Validity of the Massachusetts Candidate Assessment of Performance

Bingjie Chen
American Institutes for Research/CALDER

James Cowan
American Institutes for Research/CALDER

Dan Goldhaber
American Institutes for Research/CALDER
University of Washington

Roddy Theobald
American Institutes for Research/CALDER

Contents

Acknowledgements

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H170025 to the American Institutes for Research (AIR). All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Department of Education or the institutions to which the authors are affiliated. The authors wish to thank partners at the Massachusetts Department of Elementary and Secondary Education, including Claire Abbott, Meagan Comb, Carrie Conaway, Nicole DellaRocco, Matt Deninger, Sandra Hinderliter, Bob Lee, Liz Losee, Adrienne Murphy, Shelagh Peoples, and Heather Peske, as well as external project advisors, including Joe Berger, Joe Cambone, Clara Kang, Liesl Martin, Kathy McNamara, Heather Pacheco-Guffrey, Matt Ronfeldt, and Jim Wyckoff, for comments that improved this analysis.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
202-403-5796 • www.caldercenter.org

From the Clinical Experience to the Classroom: Assessing the Predictive Validity of the Massachusetts Candidate Assessment of Performance

Bingjie Chen, James Cowan, Dan Goldhaber, and Roddy Theobald

CALDER Working Paper No. 223-1019
October 2019

Abstract

We evaluate the predictive validity of the Massachusetts Candidate Assessment of Performance (CAP), a practice-based assessment of teaching skills that is typically taken during a candidate's student teaching placement and is a requirement for teacher preparation program completion in Massachusetts. We find that candidates' performance on the CAP predicts their in-service summative performance evaluations the following year and provides a signal of future teacher effectiveness beyond what is already captured by the state's traditional licensure tests. Our findings add to a growing literature demonstrating that it is possible to collect information about the skills of *prospective* teachers during their teacher preparation experience that are predictive of the *in-service* outcomes of teachers.

1. Introduction

One of the most pressing questions facing state education systems is how to ensure that prospective teachers have adequate teaching competence before they have classroom responsibilities of their own. While nearly every state in the country requires candidates to pass licensure tests of their basic skills and/or subject-specific knowledge as a requirement for licensure, states are increasingly adopting authentic, or performance-based, assessments that candidates must pass as an additional licensure or preparation program requirement. Yet there is relatively little evidence about whether these performance-based assessments are related to the in-service performance of teachers.[1]

Massachusetts developed and utilizes the Candidate Assessment of Performance (CAP), a practice-based assessment of teaching skills that is the centerpiece of the state's efforts to assess the quality of teacher candidates before they enter the state's teaching workforce. The CAP is typically taken during a candidate's student teaching placement and requires teachers to demonstrate evidence of effective classroom practice. Passing the CAP is high-stakes in that it became a requirement for teacher preparation program completion in Massachusetts in the 2016-17 school year.

It is important to distinguish the CAP from traditional licensure tests that prospective teachers are also required to pass. In particular, the CAP is designed to assess teaching skills that are closely aligned with the state's Standards for Effective Practice and thus provides a direct link between teacher candidates' preparation and the professional standards expected of them as

[1] The CAP is similar in concept to the widely adopted edTPA: As of 2017-18, the edTPA is offered in 41 states, and passing the edTPA is a requirement for eligibility to teach in 18. For more on the recent (and rapid) adoption of the edTPA, see Hutt, Gottleib, and Cohen (2018), and for the relationship between the edTPA and student achievement, see Goldhaber Cowan, and Theobald (2017).

Massachusetts teachers. The CAP also consists of both a formative and summative assessment (typically taken near the midpoint and end of a candidate's student teaching placement, respectively), on which candidates are evaluated on six different standards and three different dimensions of their teaching competence.[2] Thus the CAP has the potential to provide nuanced and timely feedback about the specific skills and competencies of individual candidates to the candidates themselves and their teacher preparation programs to drive candidate professional development and teacher preparation program improvement.

But for the CAP to function as conceived, the information that candidates, programs, and the state receive from the CAP should predict how candidates will perform in the state's teaching workforce. In this paper we describe research testing the ability of CAP performance to predict future in-service performance evaluations. This study builds on prior work on the predictive validity of other preservice requirements.[3] But this is among the first studies to evaluate the predictive validity of a state-developed preservice performance assessment that is *explicitly intended* to align with the evaluation process and teaching standards candidates will experience as educators in that state.

We find that candidates' performance on the CAP during the first year of statewide implementation predicts the in-service performance evaluations of those candidates who are teachers of record the following year. This is true both of the summative CAP scores that are used to determine graduation eligibility and of the formative CAP scores that are used to provide feedback to candidates during a candidate's student teaching placement. These relationships hold

[2] As described in the next section, candidates are evaluated along three dimensions ("Quality," "Scope," and "Consistency") on six different standards, or "rubric elements": Well-Structured Lessons; Adjustment to Practice; Meeting Diverse Needs; Safe Learning Environment; High Expectations; and Reflective Practice.
[3] For example, on licensure tests (Clotfelter et al., 2007; Goldhaber, 2007; Goldhaber et al., 2017b; Hendricks, 2014), the edTPA or other authentic preservice performance assessments (Bastian et al., 2016, 2018; Darling-Hammond et al., 2013; Goldhaber et al., 2017a; Wilson et al., 2010), and other aspects of teacher preparation (Boyd et al., 2009; Goldhaber et al., 2017c; Ronfeldt, 2012).

whether comparisons are made within or across teacher preparation providers or programs and in models that control for candidate scores on the state's other traditional licensure tests. These findings suggest that the CAP provides a signal of candidates' teaching that is reflected in their later performance under the state's educator evaluation system and add to a growing literature demonstrating that it is possible to collect information about the skills of *prospective* teachers during their teacher preparation experience that are predictive of their *in-service* outcomes.

2. The Candidate Assessment of Performance

Massachusetts implemented the CAP as an educator preparation program completion requirement beginning in the 2016-17 school year as a key part of its reforms to teacher evaluation and preparation.[4] Similar to other performance-based assessments like the edTPA or National Board for Professional Teaching Standards (NBPTS) portfolio assessment, it relies on multiple sources of evidence, including observations of classroom teaching practice. In addition, the CAP includes student feedback from a classroom survey and an indication of progress on some selected measure of student growth.

As noted above, the CAP is intentionally aligned with the Massachusetts Educator Evaluation system, under which teachers are evaluated according to their performance on the state's Standards for Effective Practice. The CAP evaluation cycle consists of five steps intended to mimic the steps on the state's in-service evaluation cycle.[5] First, the candidate assesses his or her own practice and identifies a potential professional development goal. Second, the candidate,

[4] The CAP was piloted to a small sample of candidates without any stakes attached in 2015-16. While most candidates take the CAP during their student teaching placement, teachers of record who are enrolled in a preparation program either to add a credential or to advance to initial certification complete the CAP while they are employed as a teacher.

[5] For more information about the state's in-service evaluation system, see Cowan et al. (2018).

program supervisor (the university faculty member who advises the teacher candidate), and supervising practitioner (the in-service teacher who supervises the candidate) meet and finalize a professional growth plan. During the third phase, the candidate works toward the professional development goal, while the supervising practitioner and program supervisor conduct observations. Fourth, at the midpoint of the evaluation cycle, the candidate receives formative feedback intended to guide further practice and professional development. Finally, at the conclusion of the evaluation cycle, the candidate receives the summative feedback that determines the final CAP outcome.

For the typical candidate who takes the CAP as part of a preservice teacher preparation program, the evaluation cycle above takes place during the candidate's student teaching practicum (and in the classroom of the candidate's supervising practitioner). The evaluation cycle is similar for teachers of record who are enrolled in a preparation program to advance certification (these comprise 25% of the sample of CAP participants), though this evaluation cycle occurs in the teacher's *own* classroom and can include activities related to the teacher's in-service performance evaluations that year.[6]

To illustrate how this evaluation process works in practice, we include an example CAP rubric in Figure 1. As part of both the formative and summative feedback, candidates are evaluated on six sub-standards from the state's Standards for Effective Practice that were judged by Massachusetts as being necessary for teacher success on Day 1 in the classroom and thus

[6] Specifically, state guidelines for the CAP state that "Candidates that are employed as teachers-of-record are still required to undergo CAP for program completion. Candidates and Sponsoring Organizations may leverage activities associated with in-service evaluations to support CAP and reduce duplication of efforts, but evaluation ratings provided by a school/district evaluator may not replace or substitute for CAP ratings. Proficiency on one does not necessitate proficiency on the other" (Massachusetts Department of Elementary and Secondary Education, 2016, p. 6). Most teachers of record are working on a provisional teaching license, which permits teachers who have passed the Massachusetts teacher licensure tests to work in public schools for up to 5 years before advancing to an initial teaching license.

comprise the CAP "rubric elements" (see panel A of Figure 2 for all of these rubric elements or sub-standards). For each of these rubric elements, candidates can receive scores of "Exemplary," "Proficient," "Needs Improvement," or "Unsatisfactory" along the three dimensions ("Quality," "Scope," and "Consistency") upon which teacher candidates are judged (see Figure 1 for formal definitions of each of these terms). Additionally, as we noted above, candidates receive both formative and summative assessments (which are based on the exact same rubric), though it is only the summative assessment that factors into passing requirements. Teacher candidates pass the CAP if they receive at least a "Proficient" rating on the "Quality" dimension *on all six rubric elements* on the summative assessment and at least a "Needs Improvement" rating on the other two dimensions for each rubric element on the summative assessment.[7] In Section 4, we describe how we create quantitative measures of CAP performance from these ordinal (but discrete) assessment scores.

As with the Massachusetts Educator Evaluation Framework, the CAP relies on the professional judgment of evaluators and permits substantial local autonomy; specifically, the responsibility for CAP scoring falls on the program supervisors and supervising practitioners themselves.[8] This sets the CAP apart from similar assessments, like the edTPA, which rely on centralized scoring by a testing company (in the case of the edTPA, Pearson). The state, however, does attempt to ensure comparability of CAP scoring through the program approval process and by offering tools and trainings to support evaluator calibration. Moreover, while Massachusetts sets the minimum standards for each domain, as described above, programs may

[7] We do find a small number of cases in which a candidate received a passing score despite not meeting the published requirements for passing the test. These are likely due to errors in local implementation of the CAP grading rubric.

[8] These roles are somewhat different for current teachers of record, who comprise 25% of the sample of CAP participants. These teachers have primarily entered teaching on a provisional teaching license and are attempting to advance their license to a standard (initial) teaching license. For these teachers, the supervising practitioner is often a mentor teacher working in the same school, and candidates complete the evaluation in their own classroom.

require higher thresholds or documentation if they choose.[9] These design decisions all reflect the state's Educator Evaluation Framework but may also lead to differences in grading standards across the state. In Section 5, we discuss our approaches to incorporating these issues into the validity analysis.

3. Prior Literature

A number of studies have examined the relationship between specific licensure tests and teacher outcomes (Clotfelter et al., 2007; Goldhaber, 2007; Goldhaber et al., 2017b; Hendricks, 2014). These tend to find modest positive relationships between teachers' licensure exam performance and teacher value-added, but the magnitudes of the estimated relationships also vary by test, grade level, and subject taught.[10] There are far fewer studies of newer performance assessments like the CAP. The earliest antecedent may be the portfolio assessment offered by the National Board for Professional Teaching Standards (NBPTS), which prior studies have shown to predict later teacher contributions to student learning (Cantrell et al., 2008; Cowan & Goldhaber, 2016).

The edTPA, which is based on the NBPTS assessment (Pecheone et al., 2013), is the most widely used performance-based assessment for preservice teacher candidates; as of 2017-18, the edTPA was offered in 41 states, and passing the edTPA was a requirement for eligibility to teach in 18 (Hutt et al., 2018). Darling-Hammond et al. (2013) found a positive relationship between a precursor of the edTPA and teacher value-added in California. More recent research has found similar relationships between the edTPA and teacher value-added in North Carolina

[9] For example, Boston College University requires the collection of additional elements not found in the CAP rubrics as part of their CAP process (Elizabeth Losee, personal communication, June 2019).
[10] For instance, Goldhaber et al. (2017b) found substantially larger relationships between science licensure test performance and teacher effectiveness in high school biology than between math licensure test performance and teacher effectiveness in secondary math.

and Washington (Bastian et al., 2016 Goldhaber et al., 2017a). For example, Goldhaber et al. (2017a) found that candidates' edTPA scores in Washington are a significant predictor of student mathematics (but not ELA) achievement in their classrooms once they enter the workforce.

Unlike traditional licensure tests, performance-based assessments like the CAP and edTPA rely on individual observers evaluating teaching practice in a classroom setting rather than a standardized assessment of content or pedagogical knowledge. Although this arguably results in a better measurement of teaching practice, researchers have also found that observers may have trouble separating teaching practice from the context in which it occurs. A number of studies, for instance, have found that teachers tend to receive higher scores on observational evaluations when they are assigned to classrooms with higher achieving students or more economically advantaged students (Campbell & Ronfeldt, 2018; Cowan et al., 2018; Gill et al., 2016; Steinberg & Garrett, 2016).

The CAP differs from the in-service observational evaluations described above in that the evaluators are the candidate's supervising practitioner and/or field supervisor rather than principals or district officials. On the one hand, local observers—especially the classroom teacher—may better understand the classroom context and adjust their ratings to account for disruptive students or other classroom factors. However, raters with personal relationships tend to provide higher scores on observational rubrics and portfolio-based certification tests (Bastian et al., 2016; Ho & Kane, 2013), particularly when there are stakes attached (Grissom & Loeb, 2017), and may provide less honest opinions than individuals without a personal connection (Leising et al., 2010).

4. Data

4.1 Candidate Assessment of Performance

For the purposes of this study, we focus on the CAP performance of 3,506 teacher candidates who took the CAP during the 2016-17 school year—the first year in which all candidates took the assessment and scores were used to determine program completion eligibility—and whose scores were provided to the state by their teacher preparation program.[11] Before providing an overview of these data, we caution that it is likely that some preparation programs did not provide CAP scores for candidates who either did fail or were likely to fail the assessment (and thus were "counseled out" of their preparation program). In fact, only 12 candidates (or 0.3% of CAP participants in the data collected by the state) whose summative CAP scores were provided to the state in 2016-17 did not pass the test, though an additional 24 candidates received scores that should not have resulted in a passing score according to minimum passing requirements established by the state—in most cases, receiving a "Needs Improvement" on at least one "Quality" dimension—yet are indicated as having passed the test.[12] Supplemental data provided by Massachusetts suggest that 138 candidates exited their program in 2016-17, which provides an upper bound for the number of teacher candidates for whom we have missing CAP performance data (i.e., at most 3–4% of all teacher candidates).[13]

Panel A of Figure 2 shows the distribution of ratings on each of the 18 scores—three dimensions for each of the six rubric elements on each assessment—for candidates who have

[11] The CAP data collected by the state also provide additional information about teacher candidates, including their program area (e.g., elementary or special education) and program type (e.g., baccalaureate or post-baccalaureate).
[12] This inconsistency is likely due to errors in local implementation of the CAP grading rubric.
[13] Note that we are not missing teacher performance data on these candidates, as they would not have been deemed eligible to teach in Massachusetts.

scores on both the CAP formative and summative assessments in 2016-17.[14] Several trends are apparent from these raw scores. First, scores tend to increase from the formative to summative assessment, as candidates are more likely to be evaluated as "Needs Improvement" on the formative assessment and "Proficient" or "Exemplary" on the summative assessment. Second, scores are generally higher on the "Quality" dimension than the "Scope" or "Consistency" dimensions, particularly on the summative assessment, which is not surprising given that a "Proficient" on all six "Quality" dimensions on the summative assessment is required for passing, while a "Needs Improvement" is sufficient on the other two dimensions. Finally, practically no candidates receive an "Unsatisfactory" rating on any of these 18 scores on either assessment, which is consistent with data on in-service teacher evaluations (Kraft & Gilmour, 2017).

As described in Section 2, Massachusetts sets minimum standards for each domain of the CAP rather than requiring that candidates surpass a particular aggregated score (as on the edTPA). However, for the purposes of this study, we aggregate the 18 scores summarized in panel A of Figure 2 into a final CAP formative score and final CAP summative score. We do this by assigning numerical values to each of the possible scores—4 for "Exemplary," 3 for "Proficient," 2 for "Needs Improvement," and 1 for "Unsatisfactory"—and adding these values across all 18 scores collected as part of each assessment.[15] The resulting final scores range from 18 (all "Unsatisfactory") to 72 (all "Exemplary") for both the formative and summative assessments.

[14] There are only 10 candidates who take the formative but not the summative assessment, but since CAP scores are not reported to the state until the end of the evaluation cycle, it is possible that more candidates drop out between the formative and summative assessments and are not observed in the data.

[15] This method of creating aggregated scores was one of two methods developed in conversations with project partners at the Massachusetts Department of Elementary and Secondary Education. We also replicate all results with a second method in which we provide double weight to the "Quality" dimension within each of the six rubric elements, and all results are qualitatively similar.

Panels B and C of Figure 2 provide an overview of the distribution of formative and summative CAP scores across all CAP participants in the 2016-17 school year (the year of CAP data we use for this study).[16] The most striking aspect of these distributions is the share of teacher candidates—22.5% of all formative CAP participants and 35.9% of all summative CAP participants—who receive a score of 54 points (the mode) on these assessments. In over 90% of these cases on both assessments, candidates received this score because they were evaluated as "Proficient" on all 18 scores.[17] This clustering of scores on a single value perhaps suggests a lack of rigor amongst some evaluators and certainly presents some challenges in relating these scores to later teacher outcomes (we return to this issue in Section 5).

The alignment of the CAP to the state's Standards for Effective Practice presents an opportunity to create and consider sub-scores on the different CAP assessments. Specifically, two of the CAP rubric elements ("Well-Structured Lessons" and "Adjustment to Practice") are aligned with Standard 1 ("Curriculum, Planning, and Assessment"); three of the CAP rubric elements ("Meeting Diverse Needs," "Safe Learning Environments," and "High Expectations") are aligned with Standard 2 ("Teaching All Students"); and the last CAP rubric element ("Reflective Practice") is aligned with Standard 4 ("Professional Culture"). We therefore create three CAP sub-scores for Standards 1, 2, and 4 (respectively) by summing only the scores from the CAP rubric elements that are aligned with each standard. We also create separate sub-scores aligned with each of the dimensions on which candidates are evaluated ("Quality," "Scope," and "Consistency") by summing scores within each dimension across the six rubric elements. We

[16] We drop candidates with multiple CAP scores.
[17] Specifically, 20.8% of formative CAP participants and 33.6% of all summative CAP participants were evaluated as "Proficient" on all 18 scores.

standardize all of these scores across all CAP participants and consider these standardized scores for the remainder of the analysis.

4.2 Summative Performance Ratings

We link the CAP data described above to measured teacher performance and other in-service teacher attributes in the 2017-18 school year, which is included in the state's Education Personnel Information Management System (EPIMS). EPIMS includes information on teacher assignments, district evaluation data, and education status.[18] For the purposes of this study, we focus on teachers in "traditional" classroom settings in which they teach at least 10 students over the course of the school year. This excludes supplemental teaching duties (e.g., any teacher who is not assigned to a classroom of students, such as special education resource teachers or supplemental English language learner instructors); this restriction permits us to estimate models that account for the demographics of a teacher's classroom. In particular, the student demographics are key to constructing the regression-adjusted performance evaluation measures described below.

EPIMS also includes teacher performance ratings collected under Massachusetts's state evaluation framework, which (like the CAP) measures performance on the state's Standards of Effective Practice. Specifically, districts evaluate teachers under the four standards and then create a final summative performance measurement based on their professional judgment of the teacher's entire practice. Importantly, there is fairly limited variation in the final summative ratings; about 85% of teachers receive a "Proficient" (3) rating in this system, which is near the median in terms of the overall concentration of evaluation rating within a single category nationally (Kraft & Gilmour, 2017). However, given the limited variation in these overall scores

[18] EPIMS does not contain a direct measure of teaching experience, so our primary measure of experience is derived from the number of years in which we observe teachers employed in EPIMS.

and prior evidence about the sensitivity of performance ratings to classroom context (Campbell & Ronfeldt, 2018; Cowan et al., 2018; Gill et al., 2016; Steinberg & Garrett, 2016), we create regression-adjusted ratings that use performance aggregated from the individual professional standards and account for differences in teaching context and consider these as our primary outcome measures.

We construct the regression-adjusted performance evaluation measures in two steps. First, in order to use the variation in teacher performance across standards, we follow Kraft et al. (2018) and fit a graded response model to the four professional standards ratings. The graded response model permits the difficulty and discrimination of each standard to differ. The difficulty of a standard describes teachers' average performance on that standard relative to the others. The discrimination of a standard indicates the strength of the relationship between unobserved teacher quality and the observed performance ratings. More discriminatory standards will tend to have greater variation in observed ratings. Formally, for standard j and rating level k, we estimate

$$\Pr\left(Y_{ij} \geq k \middle| \theta_i\right) = \frac{\exp\{a_j(\theta_i - b_{jk})\}}{1+\exp\{a_j(\theta_i - b_{jk})\}} \qquad (1)$$

where a_j is the discrimination parameter that describes the relationship between teacher performance θ_i and the rating on standard j and b_{jk} is a threshold score for rating k on standard j. We use the empirical Bayes estimates of the θ_i as the performance rating measure. We plot the item characteristic curves, which display the probabilities of the four possible ratings on each standard associated with different values of θ_i, in Figure 3.

In the second step, we adjust the performance rating scores for differences in classroom context and school evaluation standards. Prior research has found that observational measures of teacher effectiveness are sensitive to the teachers' classroom environment (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016; Whitehurst et al., 2014). In prior work, we have also

found that schools and districts in Massachusetts differ in how they award high and low performance ratings (Cowan et al., 2018). Finally, Harris et al. (2014) has found that observational ratings of teachers differ systematically across subject and grade level. We therefore estimate regressions that adjust evaluation measures for these features. To do so, we construct a data set that links teaching assignments for all teachers in Massachusetts between 2014 and 2018 to information about the class assignment and student characteristics. We then estimate variants of the following regression by OLS:

$$\hat{\theta}_{ilgst} = X_{ilgst}\beta + \alpha_l + \theta_g + \epsilon_{ilgst}. \quad (2)$$

In equation (2), X_{ilgst} is a vector of student and classroom characteristics for teacher *i* in subject *l* in grade level *g* (elementary, middle, high school) in school *s* in year *t*. The control vector includes demographic and program participation indicators, an indicator for a special education teaching assignment, an indicator for a core academic curriculum teaching assignment, and an indicator for a formative evaluation.

Table 1 summarizes the results from preferred specifications of this regression.[19] Columns 1 and 2 provide parameter estimates from a specification with district fixed effects, while columns 3 and 4 are from a specification with a school-by-grade fixed effect. In columns 2 and 4, we restrict the sample to classrooms in Grades 4 through 12 and include lagged test scores in math and ELA. The directions of the relationships in Table 1—nearly all of which suggest that teachers who teach more disadvantaged students tend to receive worse evaluations scores—are consistent with prior work on the relationship between classroom context and teacher evaluation

[19] Preferred model specifications were selected by applying teacher-switching tests outlined in Chetty et al. (2014) to the estimates of teachers' contributions to their evaluation scores from different specifications and testing whether these estimates provided unbiased estimates of out-of-sample evaluation scores. Two specifications—that include district fixed effects and school-by-grade effects—were not rejected by this approach. The correlation between the estimates adjusted with district fixed effects and the unadjusted estimates is 0.93, while the correlation between the estimates adjusted with school-by-year effects and the unadjusted estimates is 0.88.

performance (Campbell & Ronfeldt, 2018; Cowan et al., 2018; Gill et al., 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014).

We construct the regression-adjusted teacher evaluation ratings by taking the residuals from each of the specifications of equation (2) (summarized in Table 1) and averaging for each teacher and year weighted by student enrollment in each class. We then standardize the regression-adjusted performance evaluation measures and link the 2018 ratings to teacher candidates observed in the CAP data. We refer to these measures as teachers' "contribution" to their evaluation scores because they are intended to remove all sources of variation in evaluation scores outside of the teachers' control.

4.3 Summary Statistics

Table 2 provides summary statistics of the outcome measure, CAP scores, and additional candidate-level information described above for all CAP participants (n = 3,504, column 1) and CAP participants who received a summative performance rating in a traditional classroom teaching position in 2017-18 (n = 1,221).[20] The latter sample corresponds with the analytic sample in which we can consider the relationship between CAP performance in 2016-17 and summative performance ratings in 2017-18. Given the stark differences between the CAP experiences of individuals with no prior teaching experience who are taking the CAP as part of their student teaching placement and individuals who are already teaching and taking the CAP in their own classroom, we also provide separate summary statistics for these groups of candidates (candidates with no teaching experience, who are likely taking the CAP in their student teaching

[20] An additional 289 CAP participants in 2016-17 receive a summative performance rating in 2017-18 but are not employed in a teaching position with traditional teaching responsibilities, so these candidates are not part of the analytic sample for this paper.

classroom, in columns 3 and 4 and candidates with current or prior teaching experience in columns 5 and 6).

Focusing first on all CAP participants (columns 1 and 2), the summary statistics for teachers' contributions to their evaluation scores and the CAP scores themselves (panels A and B) illustrate the differences between the analytic sample and the population of all teachers (in the case of the summative performance ratings) and the population of all CAP participants (in the case of CAP scores). Specifically, given that these scores are normalized to have a mean of zero, the negative mean of the summative performance ratings in column 2 of Table 2 reflects the fact that the average teacher in the analytic sample has lower evaluation scores than the average teacher in the state. This likely reflects the fact that this sample disproportionately consists of novice teachers, who tend to receive lower performance ratings; indeed, column 4 shows that teachers in the sample with no prior experience have substantially lower evaluation scores (by about 70% of a standard deviation) than the average teacher in the state, while teachers in the sample who have taught before are closer to the average. On the other hand, the positive means across the different CAP scores in the overall sample (column 2) reflects the fact that, perhaps not surprisingly, candidates with higher CAP scores are more likely to teach in the following year.

Comparisons between candidates with no prior teaching experience (columns 3 and 4) and current or prior experience (columns 5 and 6) illustrate the stark differences in CAP performance and outcomes between these groups of candidates. Specifically, candidates with prior teaching experience receive higher CAP scores in 2016-17 and higher evaluation scores in 2017-18 than candidates who have never taught before. Because of these differences, we

consider some models in the next section that estimate the relationship between CAP scores and evaluations scores separately for these two groups of candidates.

Panels C and D of Table 2 illustrate non-random sorting into the analytic sample by candidate program area and type. For example, consistent with prior evidence on teacher workforce entry (e.g., Goldhaber et al., 2014), math candidates are more likely to appear in the analytic sample, while candidates in elementary programs are less likely.[21] Candidates from post-baccalaureate are also more likely to appear in the analytic sample than candidates from baccalaureate programs.

Panel E shows that while almost 75% of all CAP participants are not currently teachers of record and have no prior teaching experience—which reflects the fact that the CAP is typically taken as a *preservice* test in a candidate's student teaching placement—most of the other CAP participants are current teachers of record who are enrolled in a teacher preparation program either to add an additional credential or advance to an initial teaching credential from a preliminary credential.[22] The analytic sample disproportionately consists of teachers who took the CAP as a teacher of record (e.g., over 40% of the analytic sample took the CAP as a teacher of record), which is not surprising given that not all teacher candidates enter the teacher labor market and those individuals who are already teachers in 2016-17 are quite likely to be teaching in 2017-18.

Finally, most candidates in the CAP data took the Massachusetts Tests for Educator Licensure (MTEL) in communication and literacy—which consists of separate tests in reading

[21] Candidates in special education are also more likely to appear in EPIMS than other teachers, but given that our sample restrictions disproportionately drop special education teachers from the analysis, this is not reflected in the final analytic sample.

[22] The "teacher of record" program area is used by some residency programs to distinguish their candidates from traditional baccalaureate and post-baccalaureate programs. Some of these candidates do not have current or prior experience because they are serving in non-teaching roles in 2017-18.

and writing—as a requirement for their P–12 licensure in Massachusetts. We standardize these scores across all MTEL test takers and summarize scores for candidates in the various samples in panel F of Table 2. The average candidate in each sample performs higher on each MTEL test than the average test taker in the state, and candidates that enter a teaching position in 2017-18 tend to have higher MTEL scores than those who do not. While not reported in Table 2, it is notable that the correlations between the various MTEL and CAP scores considered in this analysis are quite weak (or even negative); e.g., $r = 0.04$ between the CAP summative score and the MTEL reading test, and $r = -0.04$ between the CAP summative score and the MTEL writing test.

5. Analytic Approach

Our primary analytic approach is straightforward, though we pursue a number of extensions to these basic models. Specifically, let $C_{j(t-1)}$ be a CAP score (formative, summative, or sub-score) or vector of different CAP scores for teacher j in 2016-17. We estimate a variety of models in which the outcome Y_{jt} is the contribution of teacher j to their evaluation scores in 2017-18:

$$Y_{jt} = T_{jt}\delta + C_{j(t-1)}\gamma + \epsilon_{jt} \qquad (3)$$

In equation (3), T_{jt} is a vector of teacher characteristics for teacher j in year t; as described below, the base model omits these controls, but we add specific teacher variables across other specifications. The coefficient of interest (γ) represents the expected increase in teachers' contributions to their summative performance ratings associated with a one standard deviation increase in the given CAP score.[23] Because summative performance ratings are *not* standardized

[23] We also extend the linear specification in equation (2) and model these ordinal ratings using an ordered logit model that predicts the log odds of receiving a summative performance rating of at least k ($k = 2, 3, 4$) relative to

across districts, our preferred approach uses the regression-adjusted performance measures discussed above.

While these models permit clean comparisons across different teaching contexts, they are subject to several drawbacks. First, the grading standards within providers (i.e., the institutions of higher education in which a candidate is enrolled) or programs (i.e., the specific teacher preparation program a candidate attends within a provider) could be correlated with the average effectiveness of their graduates. That is, if providers or programs producing more effective teachers have stricter standards on the CAP, then the relationship between CAP performance and teacher effectiveness will be weaker overall than it is within providers or programs. We therefore estimate all models both with and without provider and program fixed effects. Each specification has advantages and disadvantages; models without provider or program fixed effects permit comparisons across all CAP participants at the cost of potential bias due to differing CAP grading standards and aggregated outcomes across providers or programs, while models with provider or program fixed effects account for these differences at the cost of only making comparisons within providers or programs.

To explore whether different parts of the CAP provide more signal about future teacher summative performance ratings than others, we include the different CAP standard and dimension scores described in Section 4 as separate predictors in the model in equation (3). It is also of interest to examine whether a candidate's CAP performance on a given standard is more predictive of their future summative performance ratings on that standard than other standards, but we test this possibility and do not find evidence of differential predictive power across rating

receiving a summative performance rating less than k: The results from these ordered logit models tend to be very consistent with the linear models described above, so we do not discuss these estimates in our primary results.

standards. We therefore just use these CAP standards to predict the overall measures of teachers' contributions to their summative performance ratings.

We also add teacher-level control variables to the vector T_{jkt} to equation (3) to test whether the CAP predicts future performance conditional on other information about teaching effectiveness. For example, we are interested in whether the CAP provides a signal of teacher effectiveness *beyond* what is already captured by the Massachusetts Tests for Educator Licensure (MTEL), which are required for teacher licensure in the state. We therefore estimate specifications that control for candidates' scores on the two MTEL tests required of all candidates, the MTEL communication and literacy tests in reading and writing.[24]

While our primary models omit controls for teacher experience because part of the signal that the CAP provides about teacher effectiveness may be captured by differences in CAP performance by prior teaching experience (shown in Table 2), the relationship between CAP scores and in-service outcomes may be different for candidates with no prior experience than for teachers of record. We therefore estimate some models separately for teachers with and without prior teaching experience, and also pursue several additional sub-sample analyses such as limiting the sample to teachers within a given program area, program type, and career path (e.g., being hired into the same school in which the student taught).

Finally, we test the linearity assumption in the relationships between summative CAP scores, formative CAP scores, and teacher outcomes in two ways. First, we estimate flexible non-parametric local linear specifications of the model in equation (3) that allow the relationship between these two CAP scores and the outcomes to vary throughout the distribution of these scores. Second, because of our particular concern about candidates who receive a "Proficient" on

[24] Candidates are also required to pass additional subject-specific tests to receive subject-area endorsements, but we do not consider these additional tests because they are not taken by all candidates in the sample.

all 18 scores (see Figure 2), we estimate some models that identify these candidates with an additional indicator variable to test whether these candidates have systematically different outcome than other candidates (conditional on the linear relationship modeled in equation [3]).

6. Results

6.1 Results for Full Sample

Table 3 presents the estimated relationships between candidates' standardized CAP scores in 2016-17 and their standardized contributions to their summative performance ratings in 2017-18.[25] To contextualize the magnitudes of these relationships, we note that the average difference in teachers' contributions to their summative ratings in their second year of teaching relative to their first year of teaching in 0.270. The estimates in panel A are estimated across the entire sample of candidates with observed CAP scores in 2016-17 and summative performance ratings in 2017-18, and they demonstrate that CAP scores are predictive of future summative performance ratings. For example, across all candidates in the sample, a one standard deviation increase in a candidate's summative CAP score is predictive of a 0.149 standard deviation increase in the summative performance rating outcome measure the following year (column 1), which is over half of expected return to the first year of teaching to summative performance ratings.

The relationship between CAP performance and summative performance ratings is slightly (though not statistically significantly) higher for the formative CAP score (column 2), though, interestingly, only the formative CAP score is significantly predictive of summative

[25] We focus in this section on estimates of teachers' contribution to their evaluation scores estimated from a district fixed effects model but provide analogous tables based on estimates from school-by-grade fixed effects models in Appendix Tables A1–A3.

performance ratings when both the summative and formative scores are included as predictors (column 3). This may be evidence that, as has been found in some prior research on formal and informal principal evaluations (e.g., Harris & Sass, 2014; Jacob & Lefgren, 2008), scores *without* stakes attached may provide a stronger signal about teacher effectiveness. Columns 4 and 5 of panel A also show that CAP scores are still significantly predictive of future summative performance ratings when the model controls for candidate performance on the MTEL, which implies that the CAP provides a signal of future teacher effectiveness beyond what is already captured by these existing licensure tests.[26] Finally, the relationship between CAP scores and summative performance ratings is somewhat attenuated but still statistically significant when comparisons are made within specific teacher preparation providers and programs (columns 6–9), which implies that the overall relationship does not simply reflect differences in grading standards or teaching quality across different providers or programs.[27]

Panels B and C of Table 3 explore the relationships between the scores on different CAP standards or dimensions and future teacher summative performance ratings. When we consider scores aligned with the different Standards for Effective Practice (panel B), we find that the score on each individual standard is a positive and statistically significant predictor of future summative performance ratings. When we include different standards within the same model, it is clear that the overall relationship is being driven by CAP Standards 1 and 2 ("Curriculum, Planning, and Assessment" and "Teaching All Students"), which are consistently statistically significant controlling for CAP Standard 4 ("Professional Culture"). This suggests that the rubric

[26] Each of these MTEL tests is a significant predictor of summative performance ratings across the full sample of test takers with these outcomes. Results are available from the authors on request.
[27] When we test models that include a separate indicator for candidates who received a "Proficient" on all 18 scores, we find no evidence that these candidates have systematically different outcomes conditional on these linear relationships.

associated with Standard 4 provides little additional information about teacher effectiveness conditional on the other rubrics.

Likewise, a candidate's score on each of the CAP dimensions (panel C) is also a significant predictor of future summative performance ratings, and we also find evidence that these relationships are driven by scores on specific dimensions: the candidate's "Scope" and "Consistency" of teaching. This may reflect the relative importance of these dimensions or, given that candidates can pass the CAP with only a "Needs Improvement" on these dimensions, evaluators may also be using the less consequential scores to provide additional feedback to candidates on their practice.

6.2 Results for Sub-Samples

Table 4 explores the robustness of these findings to the samples described in the previous section. In panel A, we focus on teachers for whom we can calculate classroom average prior test scores and demonstrate that the relationships between CAP performance and future summative performance ratings are robust to the inclusion of these classroom controls in the model. However, relationships between CAP scores and future summative performance ratings are weaker with each group of candidates summarized in Table 4 (i.e., teachers with and without prior teaching experience) than when the two groups are combined in Table 3. While initially counterintuitive, this can be explained by the fact that teachers with prior experience receive both higher CAP scores and higher performance ratings, so a substantial portion of the overall relationship between CAP scores and summative performance ratings is driven by differences between these two groups of teachers. While this should be true by design—i.e., CAP scores should capture the fact that teachers with prior experience are more prepared for classroom responsibilities than novice educators—it does somewhat temper the predictive validity of the

CAP for novice teachers (though these relationships are still statistically significant for the formative assessment in most specifications).

We also pursue several additional sub-sample analyses by other candidate characteristics. We report one comparison in Table 5, which presents the relationships between CAP scores and summative performance ratings for candidates in the two most common program areas, elementary and special education. Despite emerging evidence (Jones et al., 2019) suggesting that *in-service* evaluation of special education teachers often does not provide a reliable and valid measure of effective special education instruction, the relationships between CAP performance and future summative performance ratings are no weaker for special education teachers than elementary teachers. While not reported in Table 5, we also find little evidence that the relationship between CAP performance and summative performance ratings differ by candidate program type (baccalaureate or post-baccalaureate). Moreover, despite there being good reason to believe that schools and districts may have more information about candidates who took the CAP in the same school or district (either because they did their student teaching there or took the CAP as a teacher of record), we do not find that the relationships between CAP scores and summative performance ratings are any stronger for these groups of teachers.[28]

The linear relationships in Tables 3–5 paint a clear picture of the predictive power of the different CAP assessments, but the existence of both a formative and summative assessment provides an opportunity to explore whether certain *combinations* of scores are more predictive of summative performance ratings than others. We therefore conclude by estimating the local linear model described in the previous section and plot predicted summative performance ratings for each observed combination of formative and summative CAP ratings in Figure 4. In Figure 4, the

[28] Results available from the authors on request.

size of each bubble indicates the number of candidates with a given combination of scores, while the color of these bubbles range from blue to red as predicted summative performance ratings increase. Specifically, the lowest predicted rating shown in Figure 4 is represented by blue and is 1.07 standard deviations lower than the average teacher in the state, and the highest predicted rating is represented by red and is 0.16 standard deviations lower than the average teacher in the state.

The primary conclusion from Figure 4 is that the pattern of results is broadly representative of the linear estimates shown in Table 3; that is, candidates with higher scores on the formative or summative CAP assessment tend to have higher predicted summative performance ratings, and there is somewhat more separation in summative performance ratings as a function of the formative assessment scores than the summative scores. This may reflect the possibility that growth from the formative to summative assessment is a separate and important indicator of future effectiveness, though we are limited in our ability to explore this possibility given that any growth measure is highly (or perfectly) collinear with the linear terms already included in the models in Tables 3-5.

7. Discussion and Conclusions

The most important conclusion of this study is that, as intended through the explicit and intentional alignment of the CAP with the Massachusetts Standards for Effective Practice, teaching candidates' scores on the CAP provide a signal of their future in-service summative performance rating beyond what is already captured by other preparation and licensure requirements in the state. This conclusion has clear implications both for Massachusetts and for other states considering performance-based assessment of teacher candidates as part of their preparation and licensure requirements. For Massachusetts, this implies that the CAP can provide

feedback about the specific skills and competencies of individual candidates to the candidates themselves and their teacher preparation programs far earlier than is typically possible with other measures of teacher effectiveness (e.g., in-service performance evaluations). And for other states, these relationships suggest that there may be advantages to state-developed assessments that align measures of candidate and teacher performance within a state.

The finding about the importance of a candidate's formative CAP score in predicting later outcomes is also novel and important. While formative assessments are a central component of many evaluation systems, the communication of these formative CAP scores *in real time* to candidates—and the distribution of these scores to preparation providers and the state as part of CAP reporting requirements—suggests that these formative assessments could play an important role in candidate professional development and teacher preparation program improvement in the state.

This study also points to potential areas of growth for CAP implementation in Massachusetts. Specifically, there at least two signs that local scoring of the CAP could be made more rigorous: the very low percentage (<1%) of CAP participants who fail the test in the data reported to the state and the significant proportion (more than a third) of candidates who are deemed to be "Proficient" on all 18 ratings. These illustrate *potential* drawbacks to state-developed assessments that rely on local scoring, but the fact that CAP scores are predictive of summative performance ratings *despite* these drawbacks suggests that there may also advantages to local implementation and scoring.

There are also two important issues not addressed by this study. The first, implicit in the theory of action associated with CAP implementation, is whether the CAP leads to improvements in the skill sets of teacher candidates (i.e., facilitates the development of teacher

candidate skills). Note that this issue is distinct from the question of predictive validity of the CAP, which is the focus of this study. In particular, there are at least two potential mechanisms through which the introduction of CAP could improve teacher candidate skills (as opposed to just providing a signal of a candidate's skills): Going through CAP could *prepare* candidates for the evaluation cycle they will experience as an in-service teacher, and the CAP could *signal* state expectations about teaching practice to candidates before they are formally evaluated. We also do not have sufficient statistical power with one year of data to evaluate the relationship between CAP performance and teachers' contributions to student learning in their classrooms (i.e., value added), which is another important dimension of teacher effectiveness, though future work will explore these relationships as well.

That said, this paper contributes to a growing literature illustrating that it is possible to learn something about the teaching skills of individual candidates during their teacher preparation experience. Unlike interventions and evaluations in the in-service teacher workforce (e.g., professional development and teacher evaluation systems), the cost of collecting this information during teacher preparation is likely lower in both monetary and political terms (i.e., because it affects teacher *candidates*, not tenured teachers). The CAP therefore represents a promising avenue for collecting this information *before* candidates have classroom responsibilities of their own and providing an opportunity to use this information for candidate development, teacher preparation program improvement, and state policy.

References

Bastian, K. C., Henry, G. T., Pan, Y., & Lys, D. (2016). Teacher candidate performance assessments: Local scoring and implications for teacher preparation program improvement. *Teaching and Teacher Education*, *59*, 1–12.

Bastian, K. C., Patterson, T. M., & Pan, Y. (2015). *UNC Teacher Quality Research: 2015 Teacher Preparation Program Effectiveness Report*. Chapel Hill, NC: Education Policy Initiative at Carolina.

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.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*. Advance online publication.

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2008). *National Board certification and teacher effectiveness: Evidence from a random assignment experiment.* 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. (2016). National Board certification and teacher effectiveness: Evidence from Washington State. *Journal of Research on Educational Effectiveness, 9*(3), 233–258.

Cowan, J., Goldhaber, D., & Theobald, R. (2018). *An exploration of sources of variation in teacher evaluation ratings across classrooms, schools, and districts* (CALDER Working Paper 140618).

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.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments (No. REL 2017–191). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources, 42*(4), 765–94.

Goldhaber, D., Cowan, J., & Theobald, R. (2017a). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education, 68*(4), 377–393.

Goldhaber, D., Gratz, T., & Theobald, R. (2017b). What's in a teacher test? Assessing the relationship between teacher licensure test scores and student secondary STEM achievement and course taking. *Economics of Education Review, 61,* 112–129.

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, 43,* 106–124.

Goldhaber, D., Krieg, J. M., & Theobald, R. (2017c). Does the match matter? Exploring whether student teaching experiences affect teacher effectiveness. *American Educational Research Journal, 54*(2), 325–359.

Grissom, J. A., & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy, 12*(3), 369–395.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added. *American Educational Research Journal, 51*(1), 73–112.

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

Hendricks, M. D. (2014). *Public schools are hemorrhaging talented teachers. Can higher salaries function as a tourniquet?* Association for Education Finance and Policy.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (Measures of Effective Teaching Project). Seattle, WA: Bill & Melinda Gates Foundation.

Hutt, E. L., Gottlieb, J., & Cohen, J. J. (2018). Diffusion in a vacuum: edTPA, legitimacy, and the rhetoric of teacher professionalization. *Teaching and Teacher Education: An International Journal of Research and Studies, 69*(1), 52–61.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–135.

Jones, N., Bell, C., Brownell, M., Qi, Y., Peyton, D., Pua, D., Fowler, M., & Holtzman, S. (2019). *Using classroom observations in the evaluation of special education teachers.* 2019 AEFP Conference Paper.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234–249.

Kraft, M. A., Papay, J. P., & Chi, O. L. (2018). *Teacher skill development: Evidence from performance ratings by principals*. Unpublished manuscript. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_papay_chi_2018_teacher_skill_development.pdf

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology, 98*(4), 668–682.

Massachusetts Department of Elementary and Secondary Education. (2016). *Guidelines for the Candidate Assessment of Performance: Assessment of Teacher Candidates.* Available at http://www.doe.mass.edu/edprep/cap/guidelines.html.

Pecheone, R., Shear, B., Whittaker, A., & Darling-Hammond, L. (2013). *2013 edTPA field test: Summary report.* Stanford, CA: Stanford Center for Assessment, Learning, and Equity.

Ronfeldt, M. (2012). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis, 34*(1), 3–26.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293–317.

Whitehurst, G. J. R., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Washington, DC: Brown Center on Education Policy, Brookings Institution.

Wilson, M., Hallam, P. J., Pecheone, R., & Moss, P. (2010). *Using student achievement test scores as evidence of external validity for indicators of teacher quality: Connecticut's Beginning Educator Support and Training Program*. Palo Alto, CA: Stanford Center for Opportunity Policy in Education.

Tables and Figures

Figure 1. Example CAP Scoring Rubric and Definitions

I.B.2: Adjustment to Practice				
	Unsatisfactory	Needs Improvement	Proficient	Exemplary
I-B-2. Adjustment to Practice	Makes few adjustments to practice based on formal and informal assessments.	May organize and analyze some assessment results but only occasionally adjusts practice or modifies future instruction based on the findings.	Organizes and analyzes results from a variety of assessments to determine progress toward intended outcomes and uses these findings to adjust practice and identify and/or implement appropriate differentiated interventions and enhancements for students.	Organizes and analyzes results from a comprehensive system of assessments to determine progress toward intended outcomes and frequently uses these findings to adjust practice and identify and/or implement appropriate differentiated interventions and enhancements for individuals and groups of students and appropriate modifications of lessons and units. Is able to model this element.

Formative Assessment				
Quality			*	
Scope		*		
Consistency		*		
Evidence:				

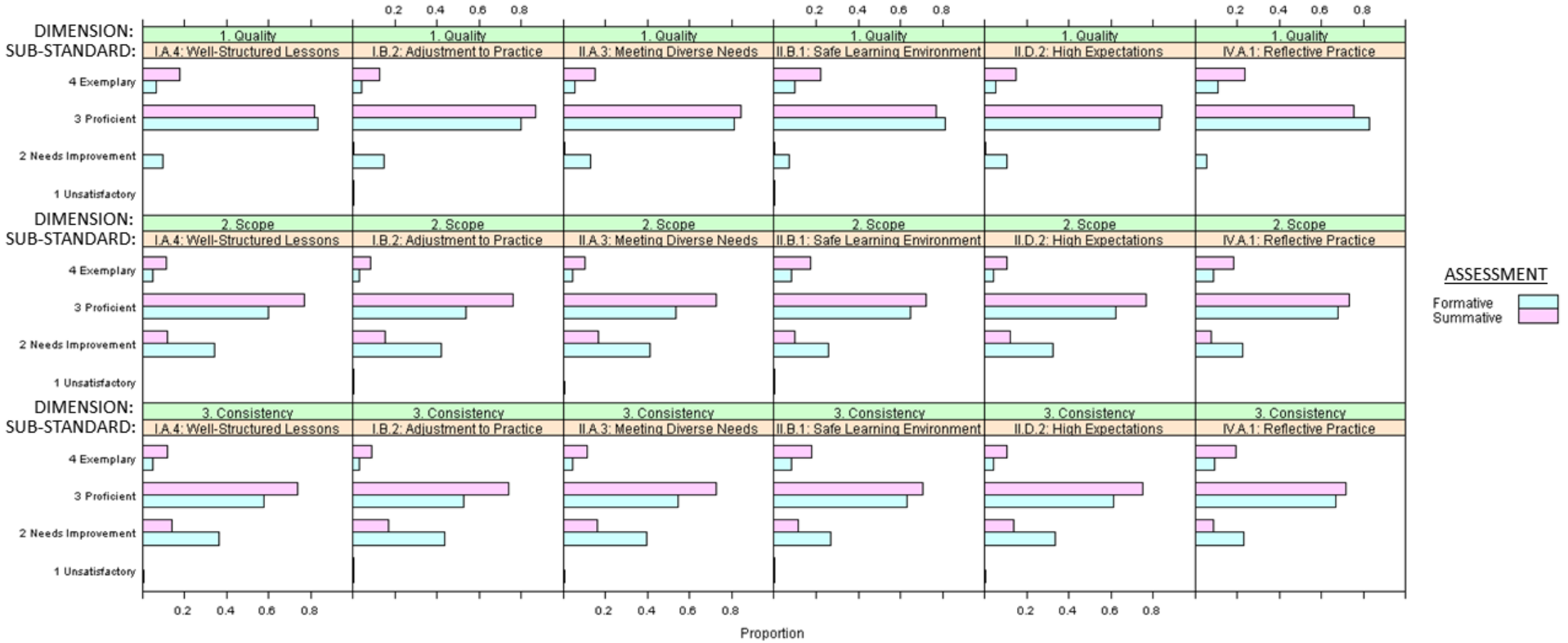
Summative Assessment				
Quality			*	
Scope		*		
Consistency		*		
Evidence:				

Additional Definitions (Massachusetts Department of Elementary and Secondary Education, 2016)

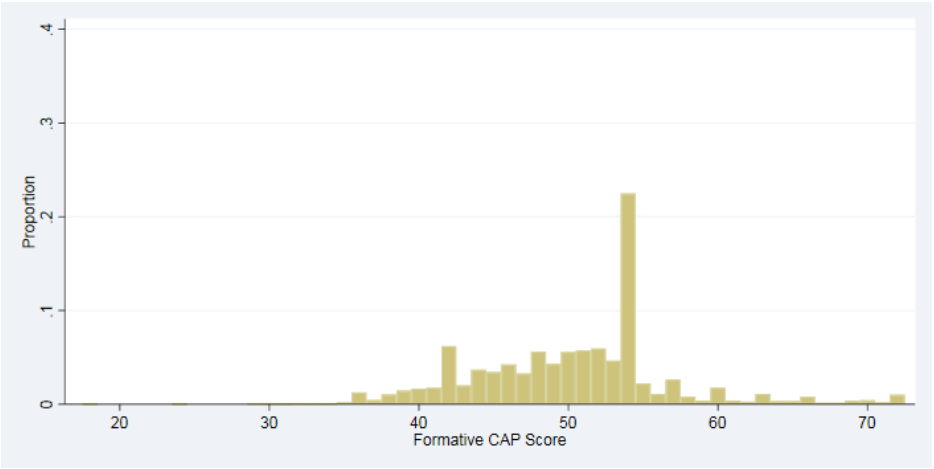
- **Quality**: the ability to perform the skill, action or behavior
- **Scope**: the scale of impact (e.g., one student, subset of children, all students) to which the skill, action or behavior is demonstrated with quality
- **Consistency**: the frequency (e.g., all the time, sometimes, once) that the skill, action or behavior is demonstrated with quality

Figure 2. Distribution of Raw and Cumulative CAP Formative and Summative Scores

Panel A. Distribution of Raw CAP Scores by Assessment, Dimension, and Sub-Standard



Panel B. CAP Cumulative Formative Scores



Panel C. CAP Cumulative Summative Scores

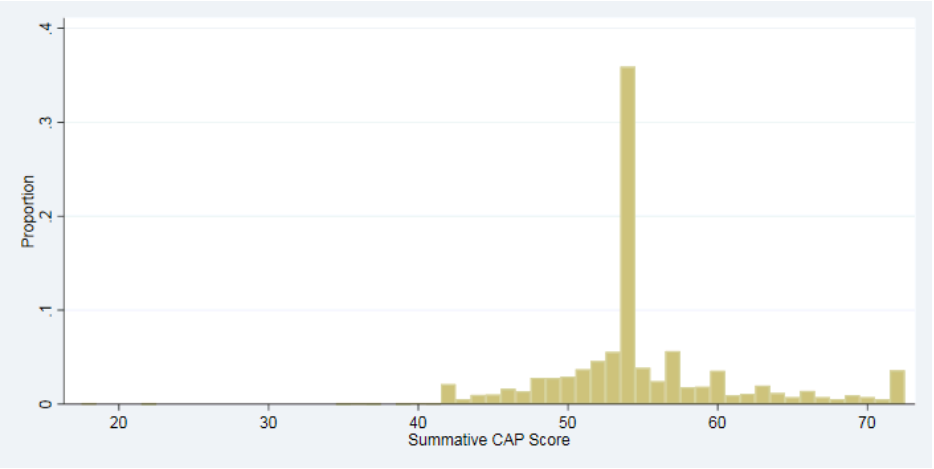
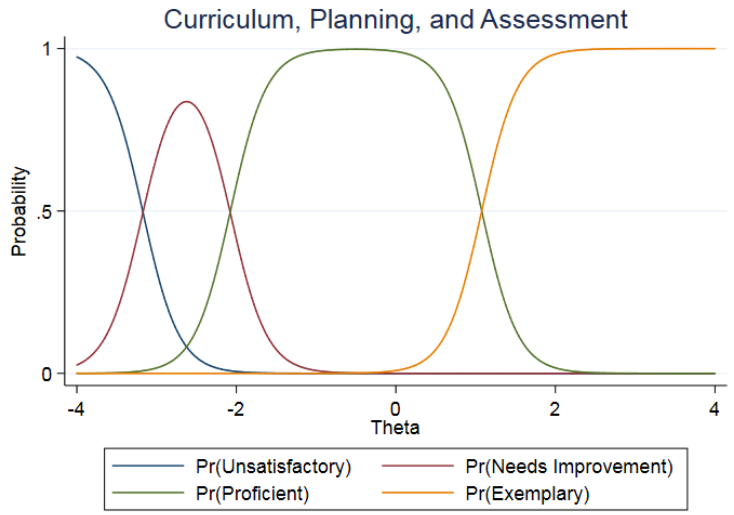
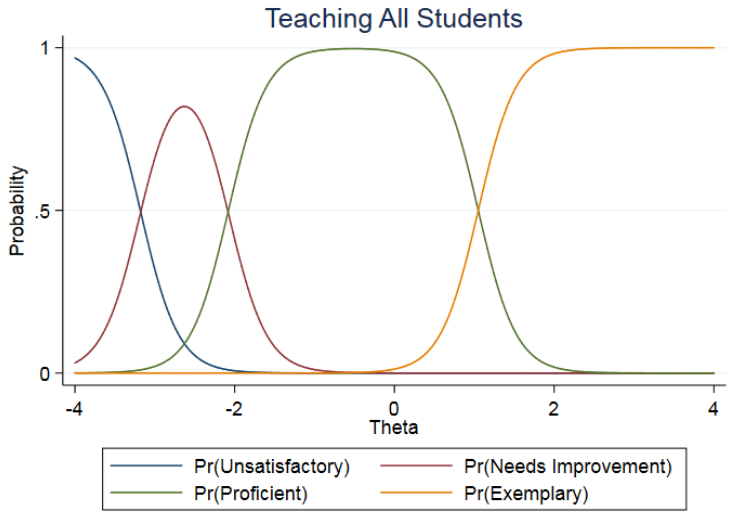


Figure 3. Item Characteristics Curves From Graded Response Model of Teacher Evaluation Scores

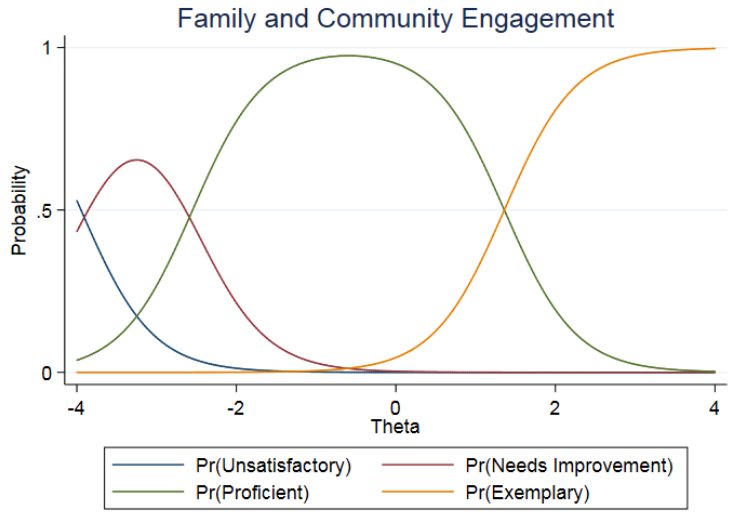
Panel A: Standard I



Panel B: Standard II



Panel C: Standard III



Panel D: Standard IV

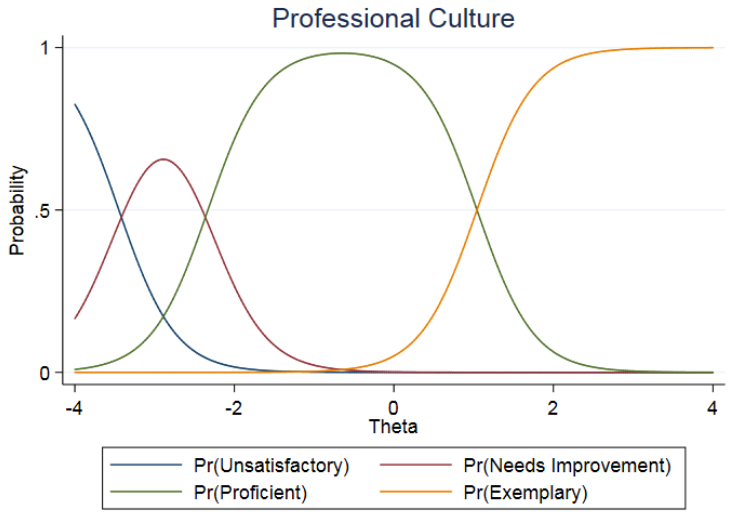
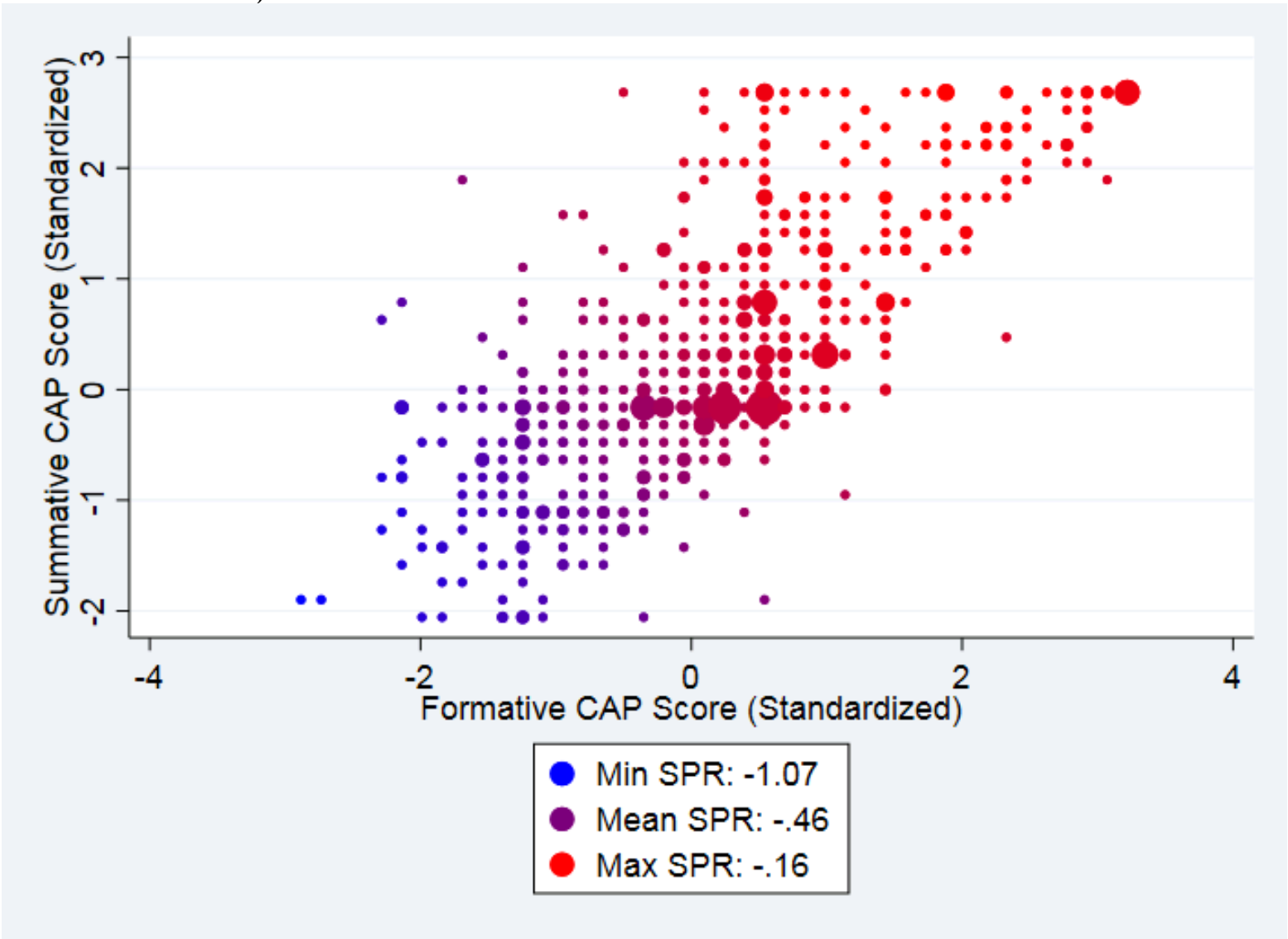


Figure 4. Predicted Teacher Contributions to Summative Performance Ratings by Formative and Summative CAP Score (District Fixed Effects Model)



Note. Size of each point proportional to number of teachers with given combination of formative and summative assessment scores.

Table 1. Regressions Predicting Aggregated Teacher Evaluation Scores

Column	(1)	(2)	(3)	(4)
Classroom proportion Limited English Proficient students	-0.069***	-0.019***	-0.009**	0.055***
	(0.004)	(0.006)	(0.004)	(0.006)
Classroom proportion economically disadvantaged students	0.137***	0.159***	-0.063***	-0.039***
	(0.003)	(0.004)	(0.004)	(0.005)
Classroom proportion male students	-0.093***	-0.083***	-0.096***	-0.091***
	(0.003)	(0.004)	(0.003)	(0.003)
Classroom proportion students receiving free/reduced priced lunch	-0.181***	-0.147***	-0.070***	-0.040***
	(0.003)	(0.004)	(0.004)	(0.005)
Classroom proportion full-inclusion students with disabilities	0.002	0.072***	-0.007*	0.041***
	(0.004)	(0.005)	(0.004)	(0.005)
Classroom proportion partial-inclusion students with disabilities	-0.077***	0.037***	-0.041***	0.046***
	(0.005)	(0.006)	(0.005)	(0.006)
Classroom proportion substantially separate students with disabilities	-0.029***	0.046***	-0.020***	0.042***
	(0.004)	(0.005)	(0.004)	(0.006)
Classroom proportion Asian students	0.117***	0.097***	0.014**	0.011
	(0.006)	(0.008)	(0.007)	(0.008)
Classroom proportion Black students	-0.143***	-0.051***	-0.078***	-0.035***
	(0.006)	(0.007)	(0.006)	(0.008)
Classroom proportion Hawaiian/ Pacific Islander students	-0.052	-0.096	-0.007	-0.042
	(0.054)	(0.064)	(0.052)	(0.061)
Classroom proportion Hispanic students	-0.094***	-0.047***	-0.047***	-0.012**
	(0.005)	(0.006)	(0.005)	(0.006)
Classroom proportion American Indian students	0.035	0.073*	-0.006	0.029
	(0.032)	(0.038)	(0.030)	(0.036)
Formative assessment	0.094***	0.085***	0.099***	0.091***
	(0.001)	(0.001)	(0.001)	(0.001)
Special education teacher assignment	-0.064***	-0.052***	-0.073***	-0.062***
	(0.003)	(0.004)	(0.003)	(0.004)
English Second Language teacher assignment	-0.013*	0.022**	-0.025***	-0.019**
	(0.007)	(0.009)	(0.007)	(0.009)
Resource teacher assignment	0.063***	0.006	0.047***	-0.010
	(0.017)	(0.020)	(0.016)	(0.019)
Class size	-0.000***	-0.001***	-0.000	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Classroom average prior ELA scores		0.038***		0.033***
		(0.003)		(0.003)
Classroom average prior math scores		0.059***		0.044***
		(0.002)		(0.003)
Subject area controls	X	X	X	X
School-level controls	X	X	X	X
District fixed effects	X	X		
School-by-grade fixed effects			X	X
Observations	1,979,789	1,438,317	1,979,789	1,438,317

Note. P-values from two-sided t-test: * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

Table 2. Summary Statistics

Candidates:	All Candidates		No Prior Experience		Prior Experience	
Column	(1)	(2)	(3)	(4)	(5)	(6)
Sample:	2016-17 CAP	2017-18 teachers	2016-17 CAP	2017-18 teachers	2016-17 CAP	2017-18 teachers
Panel A: Outcome Measures						
Teacher contribution to SPR, district FE model (std)		-0.418 (1.000)		-0.712 (0.962)		-0.067 (0.934)
Teacher contribution to SPR, school-by-grade model (std)		-0.412 (1.012)		-0.711 (0.997)		-0.078 (0.933)
Panel B: CAP Scores						
CAP Summative Score (std)	0.000 (1.000)	0.156 (0.974)	-0.090 (0.959)	-0.063 (0.855)	0.258 (1.068)	0.392 (1.072)
CAP Formative Score (std)	0.000 (1.000)	0.176 (0.999)	-0.127 (0.975)	-0.083 (0.927)	0.357 (1.086)	0.471 (1.001)
Panel C: Candidate Program Areas						
Elementary	0.226	0.172	0.274	0.257	0.088	0.076
Special Education	0.235	0.226	0.209	0.173	0.309	0.289
Early Childhood	0.095	0.077	0.110	0.093	0.050	0.063
English	0.075	0.091	0.076	0.106	0.072	0.074
Math	0.065	0.110	0.046	0.089	0.121	0.138
History	0.059	0.058	0.065	0.067	0.041	0.048
English Learners	0.035	0.043	0.020	0.024	0.078	0.063
Other	0.211	0.223	0.200	0.191	0.241	0.249
Panel D: Candidate Program Type						
Baccalaureate	0.318	0.210	0.407	0.344	0.063	0.055
Post-baccalaureate	0.566	0.677	0.500	0.553	0.756	0.820
Teacher of Record	0.063	0.067	0.036	0.052	0.139	0.090
Missing	0.053	0.047	0.057	0.050	0.042	0.035
Panel E: Candidate Teaching Experience						
No Teaching Experience	0.742	0.536	1.000	1.000	0.000	0.000
TOR With Prior Experience	0.163	0.319	0.000	0.000	0.638	0.688
Not TOR, Prior Experience	0.019	0.025	0.000	0.000	0.072	0.053
TOR, No Prior Experience	0.075	0.120	0.000	0.000	0.290	0.259
Observations (Panels A-E)_	3504	1221	2597	654	907	567
Panel F: MTEL Scores						
MTEL Communication and Literacy, Reading Score (std)	0.058 (0.889)	0.132 (0.875)	0.031 (0.900)	0.121 (0.889)	0.138 (0.850)	0.145 (0.857)
MTEL Communication and Literacy, Writing Score (std)	0.125 (0.875)	0.164 (0.906)	0.127 (0.865)	0.233 (0.888)	0.121 (0.906)	0.073 (0.926)
Observations (Panel F)	3198	1130	2407	645	791	485

Note. Standard deviations of continuous variables in parentheses. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure; SPR = Summative Performance Rating; std = standardized; TOR = Teacher of Record.

Table 3. Regressions Predicting Teacher Contributions to Summative Performance Ratings, District Fixed Effects

Column	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Summative and Formative Scores									
CAP Summative Score (standardized)	0.149*** (0.031)		0.065 (0.045)	0.144*** (0.032)		0.109*** (0.036)		0.116*** (0.039)	
CAP Formative Score (standardized)		0.161*** (0.028)	0.116*** (0.041)		0.154*** (0.029)		0.141*** (0.032)		0.130*** (0.036)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	1221	1221	1221	1130	1130	1130	1130	1061	1061
R-squared	0.021	0.026	0.028	0.020	0.025	0.082	0.089	0.170	0.173
Panel B: CAP Summative Standard-Level Ratings									
CAP Summative Standard 1	0.147*** (0.030)			0.076 (0.051)	0.159*** (0.036)		0.089* (0.052)	0.059 (0.052)	0.099* (0.057)
CAP Summative Standard 2		0.150*** (0.031)		0.087 (0.053)		0.169*** (0.038)	0.105* (0.055)	0.114** (0.054)	0.016 (0.058)
CAP Summative Standard 4			0.079*** (0.029)		-0.020 (0.035)	-0.028 (0.036)	-0.042 (0.037)	-0.024 (0.038)	0.010 (0.044)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	1221	1221	1221	1221	1221	1221	1221	1130	1061
Panel C: CAP Summative Dimension-Level Ratings									
CAP Summative Quality Dimension	0.114*** (0.029)			-0.002 (0.044)	0.017 (0.045)		-0.009 (0.046)	0.010 (0.048)	0.009 (0.055)
CAP Summative Scope Dimension		0.157*** (0.032)		0.158*** (0.050)		0.126 (0.079)	0.130 (0.081)	0.126 (0.083)	0.145 (0.093)
CAP Summative Consistency Dimension			0.149*** (0.031)		0.135*** (0.047)	0.035 (0.076)	0.038 (0.078)	0.017 (0.080)	-0.030 (0.087)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	1221	1221	1221	1221	1221	1221	1221	1130	1061

Note. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. Outcome is teacher contribution to SPR calculated from a district fixed effects model (column 1 of Table 1). P-values from two-sided t-test: * p < 0.10. ** p < 0.05. *** p < 0.01.

Table 4. Regressions Predicting Summative Performance Ratings (Standardized), District Fixed Effects Sub-Sample Models

Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Classroom Prior Performance Sample									
CAP Summative Score (standardized)	0.141*** (0.037)		0.076 (0.055)	0.135*** (0.039)		0.120*** (0.045)		0.116** (0.050)	
CAP Formative Score (standardized)		0.139*** (0.032)	0.086* (0.048)		0.136*** (0.033)		0.133*** (0.038)		0.112*** (0.043)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	892	892	892	821	821	821	821	755	755
Panel B: Teachers With No Prior Teaching Experience Sample									
CAP Summative Score (standardized)	0.064 (0.044)		0.021 (0.059)	0.066 (0.044)		0.062 (0.048)		0.056 (0.054)	
CAP Formative Score (standardized)		0.074* (0.041)	0.061 (0.055)		0.077* (0.041)		0.083* (0.043)		0.073 (0.052)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	654	654	654	645	645	645	645	591	591
Panel C: Teachers With Prior Teaching Experience Sample									
CAP Summative Score (standardized)	0.086** (0.040)		0.063 (0.062)	0.087** (0.043)		0.082* (0.046)		0.104* (0.053)	
CAP Formative Score (standardized)		0.081** (0.037)	0.033 (0.059)		0.080** (0.040)		0.101** (0.045)		0.121** (0.051)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	567	567	567	485	485	485	485	426	426

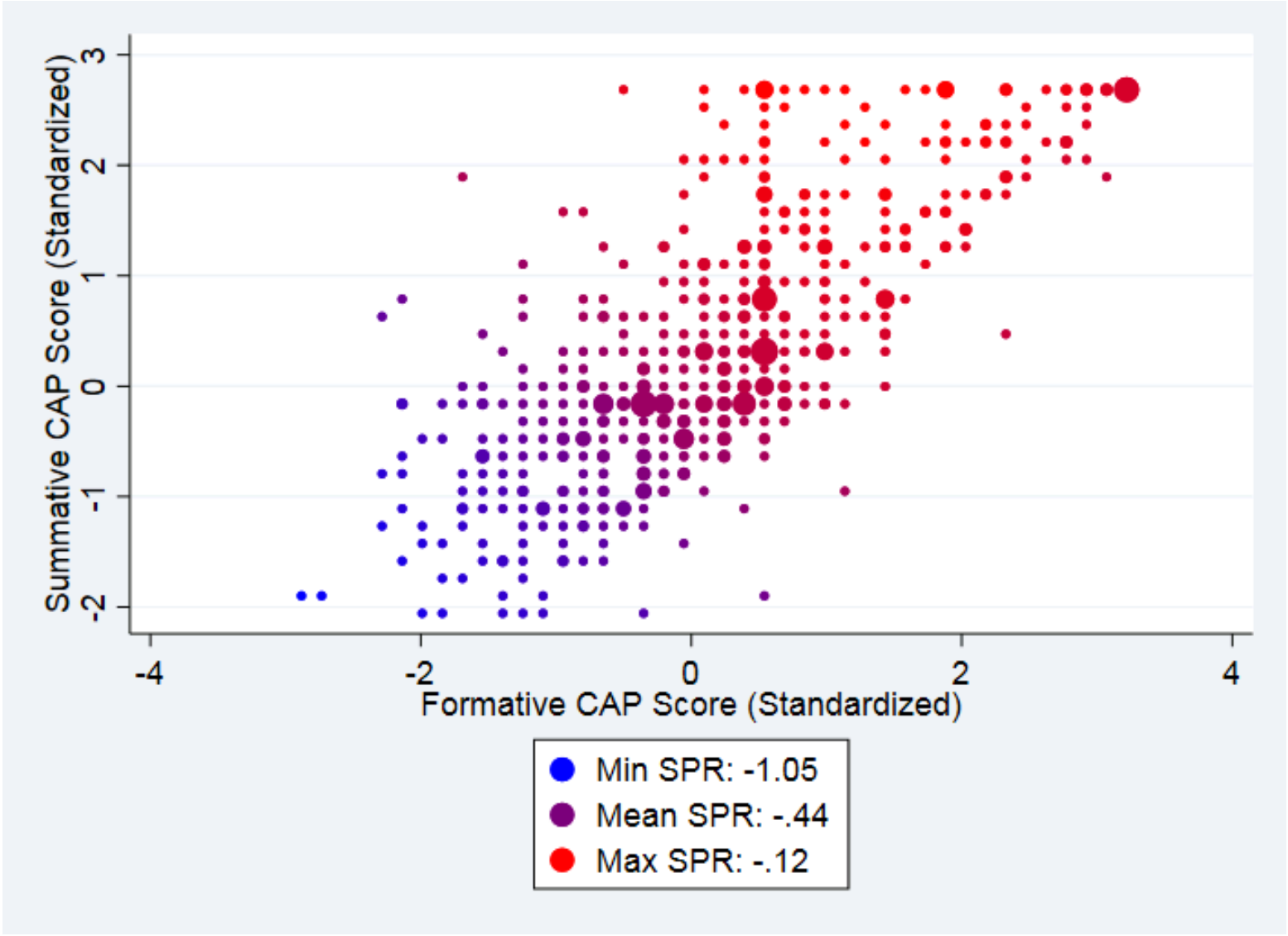
Note. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. Outcome is teacher contribution to SPR calculated from a district fixed effects model (column 1 of Table 1), including controls for classroom average prior test scores (column 2 of Table 1) in panel A only. P-values from two-sided t-test: * p < 0.10. ** p < 0.05. *** p < 0.01.

Table 5. Regressions Predicting Summative Performance Ratings (Standardized), District Fixed Effects Program Area Models

Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Elementary Program Sample									
CAP Summative Score (standardized)	0.140* (0.078)		0.112 (0.095)	0.093 (0.076)		0.072 (0.090)		0.057 (0.089)	
CAP Formative Score (standardized)		0.105 (0.080)	0.043 (0.097)		0.068 (0.081)		0.062 (0.099)		0.046 (0.098)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	216	216	216	207	207	207	207	194	194
Panel B: Special Education Program Sample									
CAP Summative Score (standardized)	0.152** (0.066)		0.154 (0.095)	0.172** (0.075)		0.096 (0.081)		0.106 (0.085)	
CAP Formative Score (standardized)		0.106** (0.054)	-0.004 (0.082)		0.112* (0.060)		0.077 (0.079)		0.096 (0.086)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	277	277	277	245	245	245	245	236	236

Note. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. Outcome is teacher contribution to SPR calculated from a district fixed effects model (column 1 of Table 1). P-values from two-sided t-test: * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

Figure A1. Predicted Teacher Contributions to Summative Performance Ratings by Formative and Summative CAP Score (School-by-Grade Fixed Effects Model)



Note. Size of each point proportional to number of teachers with given combination of formative and summative assessment scores.

Table A1. Regressions Predicting Teacher Contributions to Summative Performance Ratings, School-by-Grade Fixed Effects

Column	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Summative and Formative Scores									
CAP Summative Score (standardized)	0.151*** (0.032)		0.085* (0.045)	0.153*** (0.034)		0.120*** (0.037)		0.125*** (0.040)	
CAP Formative Score (standardized)		0.149*** (0.030)	0.090** (0.043)		0.147*** (0.031)		0.130*** (0.034)		0.117*** (0.039)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	1221	1221	1221	1130	1130	1130	1130	1061	1061
R-squared	0.021	0.021	0.025	0.021	0.021	0.083	0.086	0.171	0.171
Panel B: CAP Summative Standard-Level Ratings									
CAP Summative Standard 1	0.147*** (0.031)			0.084 (0.052)	0.146*** (0.038)		0.089* (0.053)	0.059 (0.054)	0.085 (0.058)
CAP Summative Standard 2		0.147*** (0.031)		0.077 (0.052)		0.149*** (0.039)	0.084 (0.054)	0.101* (0.055)	0.013 (0.059)
CAP Summative Standard 4			0.092*** (0.029)		0.001 (0.036)	-0.003 (0.036)	-0.017 (0.037)	0.000 (0.038)	0.039 (0.044)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	1221	1221	1221	1221	1221	1221	1221	1130	1061
Panel C: CAP Summative Dimension-Level Ratings									
CAP Summative Quality Dimension	0.112*** (0.030)			-0.006 (0.046)	-0.002 (0.047)		-0.021 (0.048)	-0.005 (0.050)	-0.009 (0.059)
CAP Summative Scope Dimension		0.155*** (0.033)		0.160*** (0.052)		0.087 (0.079)	0.097 (0.081)	0.097 (0.083)	0.112 (0.092)
CAP Summative Consistency Dimension			0.156*** (0.032)		0.158*** (0.051)	0.077 (0.077)	0.086 (0.081)	0.070 (0.083)	0.030 (0.091)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	1221	1221	1221	1221	1221	1221	1221	1130	1061

Note. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. Outcome is teacher contribution to SPR calculated from a school-by-grade fixed effects model (column 3 of Table 1). P-values from two-sided t-test: * p < 0.10. ** p < 0.05. *** p < 0.01.

Table A2. Regressions Predicting Summative Performance Ratings (Standardized), School-by-Grade Fixed Effects Sub-Sample Models

Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Classroom Prior Performance Sample									
CAP Summative Score (standardized)	0.148*** (0.038)		0.095* (0.056)	0.151*** (0.040)		0.139*** (0.046)		0.146*** (0.050)	
CAP Formative Score (standardized)		0.136*** (0.034)	0.069 (0.051)		0.141*** (0.035)		0.134*** (0.039)		0.121*** (0.045)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	892	892	892	821	821	821	821	755	755
Panel B: Teachers With No Prior Teaching Experience Sample									
CAP Summative Score (standardized)	0.063 (0.049)		0.038 (0.063)	0.066 (0.049)		0.062 (0.053)		0.057 (0.060)	
CAP Formative Score (standardized)		0.058 (0.045)	0.035 (0.059)		0.062 (0.045)		0.065 (0.048)		0.061 (0.058)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	654	654	654	645	645	645	645	591	591
Panel C: Teachers With Prior Teaching Experience Sample									
CAP Summative Score (standardized)	0.087** (0.038)		0.086 (0.059)	0.101** (0.042)		0.100** (0.046)		0.134*** (0.051)	
CAP Formative Score (standardized)		0.067* (0.038)	0.001 (0.059)		0.078* (0.041)		0.102** (0.046)		0.129** (0.052)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	567	567	567	485	485	485	485	426	426

Note. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. Outcome is teacher contribution to SPR calculated from a school-by-grade fixed effects model (column 3 of Table 1), including controls for classroom average prior test scores (column 4 of Table 1) in panel A only. P-values from two-sided t-test: * p < 0.10. ** p < 0.05. *** p < 0.01.

Table A3. Regressions Predicting Summative Performance Ratings (Standardized), School-by-Grade Fixed Effects Program Area Models

Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Elementary Program Sample									
CAP Summative Score (standardized)	0.126 (0.078)		0.124 (0.094)	0.087 (0.078)		0.087 (0.090)		0.072 (0.089)	
CAP Formative Score (standardized)		0.071 (0.081)	0.002 (0.098)		0.037 (0.084)		0.018 (0.102)		0.000 (0.103)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	216	216	216	207	207	207	207	194	194
Panel B: Special Education Program Sample									
CAP Summative Score (standardized)	0.140** (0.069)		0.154 (0.099)	0.169** (0.079)		0.100 (0.086)		0.115 (0.089)	
CAP Formative Score (standardized)		0.087 (0.060)	-0.023 (0.090)		0.098 (0.066)		0.054 (0.083)		0.083 (0.087)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	277	277	277	245	245	245	245	236	236

Note. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. Outcome is teacher contribution to SPR calculated from a school-by-grade fixed effects model (column 3 of Table 1). P-values from two-sided t-test: * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.