CALDER

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. 223-1019-2 • February 2021

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

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). 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 current and former partners at the Massachusetts Department of Elementary and Secondary Education, including Claire Abbott, Meagan Comb, Carrie Conaway, Nicole DellaRocco, Matt Deninger, Bob Lee, Liz Losee, Elana McDermott, Adrienne Murphy, Shelagh Peoples, Heather Peske, Sandra Sarucia, and Aubree Webb, 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 working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
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. 223-1019-2
February 2021

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 in their first 2 years in the teaching workforce and provides a signal of these ratings beyond what is already captured by the state's traditional licensure tests, but is not predictive of their value added to student test scores. These findings suggest that the CAP captures aspects of candidates' skills and competencies that are better reflected in their future performance evaluations than by their impacts on student performance.

1. Introduction

One of the most pressing questions faced by 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, such as the Massachusetts Tests for Educator Licensure (MTEL), which are a licensure requirement in the state. 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

[1] The CAP is similar in concept to the widely adopted Educative Teacher Performance Assessment (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 states. 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).

standards expected of them as Massachusetts teachers. The CAP also consists of both a formative and a summative assessment (typically taken near the midpoint and end of a candidate's student teaching placement) 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 Massachusetts' teaching workforce. In this paper, we describe research testing the ability of CAP performance to predict future in-service performance evaluations and value added to student test score gains. 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 that candidates will experience as educators in that state.

We find that candidates' performance on the CAP during the first 2 years of statewide implementation significantly predicts the in-service performance evaluations of candidates once they enter the state's teaching workforce. These relationships hold 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. However, we find that CAP

[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; Cowan et al., 2020; 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).

scores are significantly more predictive of candidates' summative performance ratings than their value added to student test performance—and, in fact, that CAP scores are not significantly predictive of candidates' future value added. These findings suggest that the CAP provides a better signal of candidates' teaching as reflected in their later performance under the state's educator evaluation system than in their contributions to student test score gains.

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 Educative Teacher Performance Assessment (edTPA) or National Board for Professional Teaching Standards (NBPTS) portfolio assessment, the CAP 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 in 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 and colleagues (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 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 necessary for teacher success on Day 1 in the classroom and thus comprise

[6] Specifically, state guidelines for the CAP state that "[c]andidates 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.

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 only the summative assessment 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] Most importantly for this study, recent work from Massachusetts (Cowan et al., 2020) finds that candidate scores on the state's traditional licensure tests—the MTEL, discussed in Section 1—are significantly predictive of both their summative performance ratings and their value added to student test score gains once candidates enter the state's public teaching workforce.

There are far fewer studies of newer performance assessments like the CAP. The earliest antecedent may be the portfolio assessment offered by the 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;

[9] For example, Boston College 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 and colleagues (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.

as of 2017–18, the edTPA was offered in 41 states, and passing the edTPA was a requirement for eligibility to teach in 18 states (Hutt et al., 2018). Darling-Hammond and colleagues (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 and Washington (Bastian et al., 2016, Goldhaber et al., 2017a). For example, Goldhaber and colleagues (2017a) found that candidates' edTPA scores in Washington are a significant predictor of student math (but not English language arts [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 through 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). There is also some recent evidence that variation in clinical teaching observation scores in one large university in Texas largely reflects differences in rating standards between different field supervisors rather than true differences between teaching candidates (Bartanen & Kwok, 2021).

4. Data

4.1 Candidate Assessment of Performance

For the purposes of this study, we focus on the CAP performance of 6,814 teacher candidates who took the CAP during the 2016–17 and 2017–18 school years—the first 2 years 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). For example, only 12 candidates (or 0.3% of CAP participants in the data collected by the state in 2016–17) 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, those candidates received a score of "Needs Improvement" on at least one "Quality" dimension—yet

[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).

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% to 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 scores on both the CAP formative and summative assessments.[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

[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.

[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.

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.

Panels B and C of Figure 2 provide an overview of the distribution of formative and summative CAP scores across all CAP participants.[16] The most striking aspect of these distributions is the share of teacher candidates—22.1% of all formative CAP participants and 35.7% of all summative CAP participants—who receive a score of 54 points (the mode) on these assessments. In more than 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 among 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”)

[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.
[16] We dropped 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.

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

4.2 Teacher Effectiveness Measures

We link the CAP data described above to teacher performance measures and other in-service teacher attributes in the 2017–18 and 2018–19 school years, 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 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 measures described next.

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

[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.

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" rating in this system, which is near the median in terms of the overall concentration of evaluation ratings within a single category nationally (Kraft & Gilmour, 2017). However, given the limited variation in these overall scores 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.

In order to use the variation in teacher performance across standards, we follow Kraft and colleagues (2020) 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 θ_i as the performance rating measure.

The EPIMS data can be further linked using the state's Student Information Management System (SIMS) to the demographics and test scores of the students in these teachers' classrooms. We use the performance ratings measures derived from equation (1) and student test scores from SIMS to generate the outcome measures in this study. Importantly, we adjust both measures for differences in classroom context, and in the case of performance ratings, we additionally adjust for potential differences in school evaluation standards. This is important because 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), and prior work on teacher value added has shown that models that control for these variables produce estimates of teacher contributions to student learning gains with limited bias (Bacher-Hicks et al., 2019; Chetty et al., 2014).[19]

Specifically, we construct a data set that links teaching assignments for all teachers in Massachusetts between 2014 and 2019 to information about the class assignment and student characteristics and test scores. We then estimate variants of the following models that regress student achievement or performance ratings θ_{ijt} on student controls X_{ijt}:

$$\theta_{ijt} = X_{ijt}\gamma + \epsilon_{ijt} \qquad (2)$$

In the model in equation (2), 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 Partnership for

[19] 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), and Harris and colleagues (2014) found that observational ratings of teachers differ systematically across subject and grade level.

Assessment of Readiness for College and Careers (PARCC) and online PARCC assessments.[20] In models involving teacher evaluations, we also include an indicator for a formative assessment, interact grade fixed effects with course subject, and include school fixed effects to account for differences in evaluation standards across schools (Cowan et al., 2018). We then average residuals from this regression by teacher and year to construct the measures of teacher effectiveness associated with each outcome. We refer to these measures as teachers' "contribution" to their evaluation scores or their students' test score gains because they are intended to remove all sources of variation in these measures outside of the teachers' control.

4.3 Summary Statistics

Table 1 provides summary statistics of the outcome measures, CAP scores, and additional candidate-level information described above for all CAP participants (*n* = 6,814, column 1); CAP participants who received a summative performance rating in a traditional classroom teaching position in 2017–18 or 2018–19 (*n* = 3,040); and CAP participants linked to their value added to student test score gains in 2017–18 or 2018–19 (*n* = 1,420). 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 just for candidates who did not take the CAP as a teacher of record (columns 4–6). As described in the next section, the samples in columns 5 and 6 are the analytic samples for the primary analyses in this paper—i.e., the analyses relating CAP scores to teacher effectiveness for candidates with no

[20] The structural transition control is an indicator of 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).

prior teaching experience—but we also use the samples in columns 2 and 3 for extensions in which we consider all CAP participants.

Focusing first on all CAP participants (columns 1–3), 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 1 reflects the fact that the average teacher in the analytic sample has lower evaluation scores than the average teacher in the state, even controlling for teacher experience. 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 all candidates (columns 1–3) and those with no prior experience (columns 4–6) illustrate the stark differences in CAP performance and outcomes between these groups of candidates. Specifically, candidates with no prior teaching experience receive lower CAP scores and lower evaluation scores than candidates who have prior teaching experience. Panels C and D of Table 1 also illustrate nonrandom sorting into the analytic samples 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 to appear in the sample.[21] Post-baccalaureate candidates are also more likely to appear in the analytic sample than candidates from baccalaureate programs.

Panel E shows that while more than 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 samples in columns 2 and 3 disproportionately consist of teachers who took the CAP as a teacher of record (e.g., almost 40% of the sample in column 2 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 when they participated in CAP are quite likely to be teaching in the following year.

Finally, most candidates in the CAP data took the MTEL in communication and literacy—which consists of separate tests in reading and writing—as a requirement for their P–12 licensure in Massachusetts.[23] We standardize these scores across all MTEL test takers and summarize scores for candidates in the various samples in panel F of Table 1. The average candidate in each sample performs higher on each MTEL test than the average test taker in the state, and candidates who enter a teaching position the following year tend to have higher MTEL scores than those who do not. While not reported in Table 2, it is notable that the correlations

[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 samples.
[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 the following year.
[23] Candidates are also required to take subject-area tests that we do not consider in this analysis because they are often taken after or contemporaneously to CAP.

between the various MTEL and CAP scores considered in this analysis are quite weak; e.g., $r = 0.03$ between the CAP summative score and the MTEL reading test, and $r = 0.01$ 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 be a CAP score (formative, summative, or sub-score) or vector of different CAP scores for teacher j. We estimate a variety of models in which the outcome Y_{jt} is the contribution of teacher j to their evaluation scores or student test score gains:

$$Y_{jt} = T_{jt}\delta + C_j\gamma + \epsilon_{jt} \qquad (3)$$

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. In the case of value added, we stack teacher value added across math and ELA and include a subject indicator in the model in equation (3). The coefficient of interest (γ) represents the expected increase in teachers' contributions to their summative performance ratings or student test score gains associated with a one standard deviation increase in the given CAP score.[24]

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

[24] 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 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.

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 on other standards, but we test this possibility and do not find evidence of differential predictive power across rating standards. We therefore just use these CAP standards to predict the overall measures of teachers' contributions to their summative performance ratings and student test score gains.

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 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.[25]

Finally, because of the differences between teachers of record and teachers with no prior experience illustrated in Table 1—and because the CAP is *typically* taken as a preservice assessment—we focus our prior results on candidates who had no prior teaching experience when they took the CAP (columns 4–6 of Table 1). This is analogous to prior work on traditional teacher licensure tests (e.g., Cowan et al., 2020) and the edTPA (e.g., Goldhaber et al., 2017a) that considers preservice licensure test performance and predictors of in-service teacher effectiveness. That said, because almost a quarter of CAP participants were teachers of record, we also consider extensions that include these candidates in the analysis (i.e., columns 1–3 of Table 1).

6. Results

Results for Novice Teachers

Table 2 presents the estimated relationships between candidates' standardized CAP scores and their standardized contributions to their summative performance ratings after they enter the state's public teaching workforce for the first time. 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 is 0.270. The estimates in panel A are estimated across all the candidates in column 5 of Table 1 and

[25] 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.

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 0.077 standard deviation increase in the summative performance rating outcome measure (column 1), which is over a quarter of expected increase in in summative performance ratings from teachers' first to second years in the workforce.

The relationship between CAP performance and summative performance ratings is slightly (though not statistically significantly) lower for the formative CAP score (column 2), and only the summative CAP score is significantly predictive of summative performance ratings when both the summative and formative scores are included as predictors (column 3). 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 for the CAP summative scores 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

[26] Each of these MTEL tests is a significant predictor of summative performance ratings across the full sample of test takers with these outcomes (Cowan et al., 2020).

[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.

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, the overall relationship appears to be driven by CAP Standard 1 ("Curriculum, Planning, and Assessment").

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.

In stark contrast to the results in Table 2, the results in Table 3 suggest little relationship between CAP scores—regardless of how these scores are defined across the panels of Table 3 or how the comparisons are made across the different columns of Table 3—and novice teachers' value added. In fact, not only are none of these estimates statistically significant from zero, but also the precision of the estimates in Tables 2 and 3 allow us to conclude that CAP scores are a *significantly weaker predictor* of novice teachers' contributions to student test score gains than their contributions to their summative performance ratings. While perhaps not surprising given the close and intentional alignment between the CAP and the standards used to evaluate in-service teachers in Massachusetts, this finding has important implications, which are discussed in the last section.

Because we are using multiple years of CAP data from the first 2 years of full CAP implementation, we test whether there are differences in the relationships between CAP performance and teacher effectiveness by CAP year in Table 4. There is little evidence that the relationship between CAP scores and teachers' contributions to their summative performance ratings differs for the two cohorts of CAP participants, while there is some evidence that the relationship between CAP scores and teachers' contributions to student test score gains is more positive for the 2018 CAP cohort than for the 2017 CAP cohort.

Results for All Candidates

Table 5 expands the sample to include all CAP participants—i.e., the samples in columns 2 and 3 of Table 1—and, in the even columns, includes interactions between CAP scores and an indicator for whether candidates were a teacher of record when they participated in CAP. These interactions test whether the relationships between CAP scores and teacher effectiveness vary for these teachers of record compared to the novice teachers who are the focus of Tables 2 through 4. There is little evidence that the relationship between CAP scores and teachers' contributions to their summative performance ratings differs for teachers of record relative to novice teachers, while there is only some evidence that the relationship between CAP scores and teachers' contributions to student test score gains is more positive for teachers of record than for novice teachers. This is perhaps not surprising given that, for teachers of record, the CAP scores are based on their actual in-service teaching rather than their performance in a student teaching placement.

7. Discussion and Conclusions

One conclusion of this study is that, as intended through the explicit and purposeful 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 that CAP scores are better predictors of summative performance ratings than value added is also novel and important. On the one hand, this is not surprising given the intentional alignment between CAP and the standards used to evaluate teachers in Massachusetts, but it also suggests that the CAP captures aspects of candidate skills and competencies that are better reflected in their future performance evaluations than by their impacts on student performance. This is consistent with prior work on both preservice (Bartanen & Kwok, 2021) and in-service (Cowan et al., 2018) teacher evaluations showing that these evaluations may pick up preferences for specific teaching skills across teacher education programs and schools that are not strongly correlated with student achievement gains. Combined with recent work on MTEL in Massachusetts (Cowan et al., 2020) showing that MTEL scores are significantly predictive of both summative performance ratings and value added—but less predictive of performance ratings than CAP—this suggests that the state should consider CAP and MTEL in tandem because they appear to capture different dimensions of teacher effectiveness (and are nearly uncorrelated with each other).

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 be advantages to local implementation and scoring.

At least one important issue is not addressed by this study. Specifically, 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.

That said, this paper contributes to a growing body of 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 for providing an opportunity to use this information for candidate development, teacher preparation program improvement, and state policy.

References

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review, 73*. Retrieved from https://doi.org/10.1016/j.econedurev.2019.101919

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.

Bartanen, B., & Kwok, A. (2021). Examining clinical teaching observation scores as a measure of preservice teacher quality. *American Educational Research Journal.* Retrieved from https://doi.org/10.3102/0002831221990359

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, 55*(6), 1233–1267. Retrieved from https://eric.ed.gov/?id=EJ1196784

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).

Cowan, J., Goldhaber, D., Jin, Z., & Theobald R. (2020). *Teacher licensure tests: Barrier or predictive tool?* (CALDER Working Paper No. 245-1020).

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.

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.

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. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315-347.

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.* Retrieved from 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

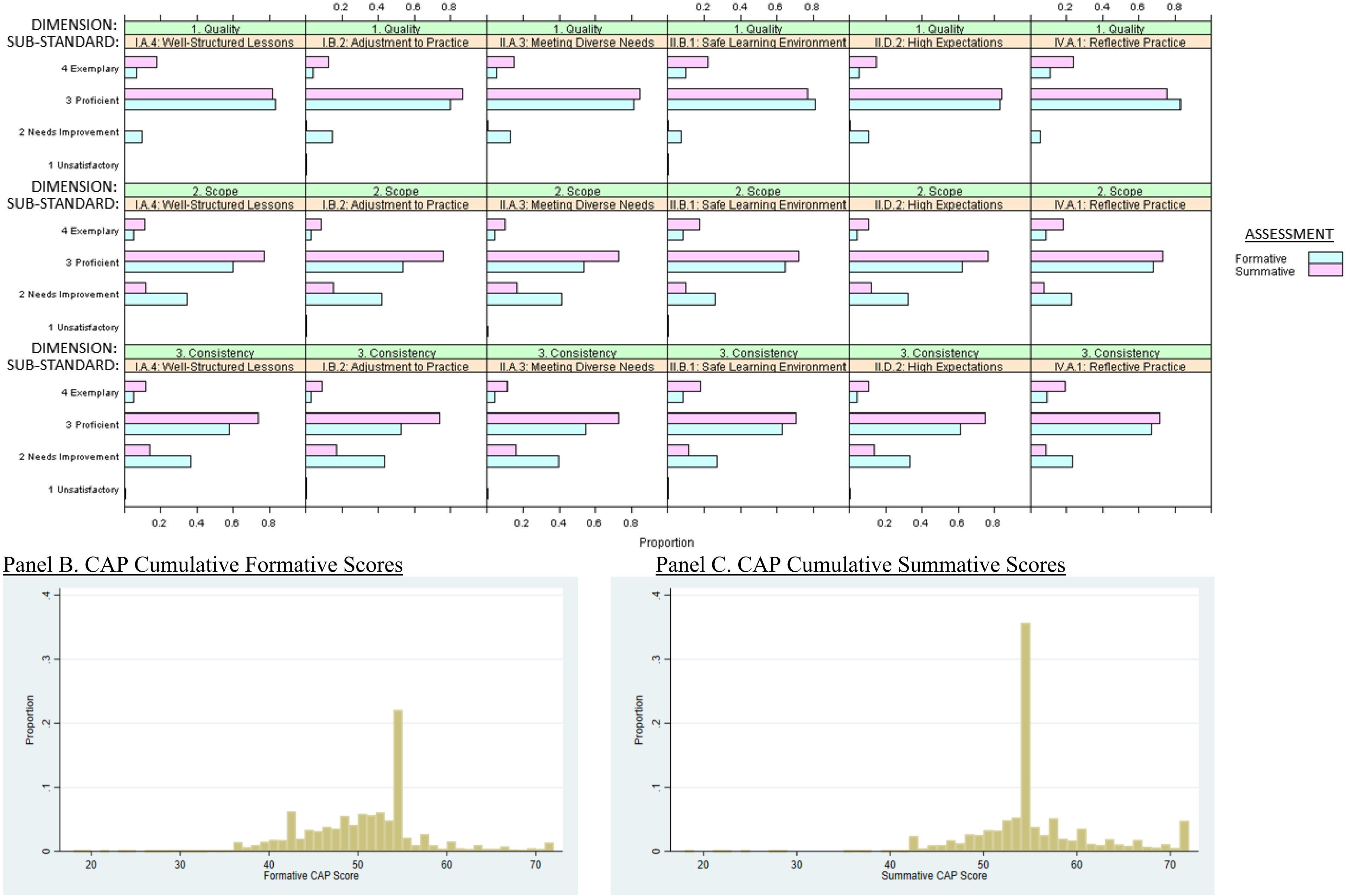
Figure 1. Example CAP Scoring Rubric and Definitions

I.A.1: Subject Matter Knowledge				
	Unsatisfactory	Needs Improvement	Proficient	Exemplary
I-A-1. Subject Matter Knowledge	Demonstrates limited knowledge of the subject matter and/or its pedagogy; relies heavily on textbooks or resources for development of the factual content. Rarely engages students in learning experiences focused on complex knowledge or subject-specific skills and vocabulary.	Demonstrates factual knowledge of subject matter and the pedagogy it requires by sometimes engaging students in learning experiences that enable them to acquire complex knowledge and subject-specific skills and vocabulary.	Demonstrates sound knowledge and understanding of the subject matter and the pedagogy it requires by consistently engaging students in learning experiences that enable them to acquire complex knowledge and subject-specific skills and vocabulary, such that they are able to make and assess evidence-based claims and arguments.	Demonstrates expertise in subject matter and the pedagogy it requires by consistently engaging all students in learning experiences that enable them to acquire, synthesize, and apply complex knowledge and subject-specific skills and vocabulary, such that they are able to make and assess evidence-based claims and arguments. Models this practice for others.
Quality			*	
Scope		*		
Consistency		*		

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

Table 1. Summary Statistics

Column	(1)	(2)	(3)	(4)	(5)	(6)
Sample:	**All Candidates**	**SPR (All)**	**VA (All)**	**All non-TOR**	**SPR (Non-TOR)**	**VA (Non-TOR)**
Panel A: Outcome Measures						
SPR		-0.083 (.723)			-0.113 (.733)	
Value Added (VA)			0.000 (.230)			-0.002 (.230)
Proportion of Math VA			0.516			0.506
Proportion of 2019 Outcome		0.668	0.658		0.675	0.667
Panel B: CAP Scores						
Proportion 2017 CAP	0.531	0.713	0.712	0.539	0.738	0.729
CAP Summative Score (std)	0.000 (.998)	0.132 (.976)	0.094 (.906)	-0.074 (.970)	-0.029 (.892)	-0.063 (.754)
CAP Formative Score (std)	-0.003 (.997)	0.160 (.979)	0.130 (.947)	-0.108 (.987)	-0.048 (.913)	-0.033 (.850)
Panel C: Candidate Program Areas						
Elementary	0.228	0.237	0.466	0.271	0.322	0.588
Special Education	0.224	0.161	0.227	0.206	0.123	0.180
Early Childhood	0.104	0.097	0.003	0.118	0.112	0.004
English	0.070	0.100	0.108	0.068	0.100	0.095
Math	0.066	0.126	0.516	0.048	0.105	0.506
History	0.056	0.070	0.011	0.058	0.074	0.010
English Learners	0.033	0.043	0.020	0.021	0.028	0.013
Other	0.219	0.167	0.064	0.210	0.136	0.050
Panel D: Candidate Program Type						
Baccalaureate	0.332	0.245	0.301	0.411	0.356	0.399
Post-baccalaureate	0.563	0.632	0.594	0.515	0.537	0.526
Teacher of Record (TOR)	0.067	0.079	0.075	0.032	0.057	0.047
Missing	0.038	0.044	0.030	0.042	0.050	0.028
Panel E: Candidate Teaching Experience						
No Teaching Experience	0.765	0.621	0.706	1.000	1.000	1.000
TOR With Prior Experience	0.141	0.263	0.199	0.000	0.000	0.000
Not TOR, Prior Experience	0.019	0.016	0.015	0.000	0.000	0.000
TOR, No Prior Experience	0.074	0.100	0.080	0.000	0.000	0.000
Observations (Panels A–E)_	6,814	3,040	1,420	5,213	1,887	1,003
Panel F: MTEL Scores						
MTEL Communication and Literacy, Reading Score (std)	0.054 (.885)	0.143 (.875)	0.148 (.812)	0.017 (.892)	0.118 (.890)	0.105 (.827)
Observations	6,017	2,762	1,304	4,607	1,774	954
MTEL Communication and Literacy, Writing Score (std)	0.119 (.848)	0.196 (.858)	0.204 (.825)	0.092 (.845)	0.198 (.854)	0.159 (.845)
Observations	6,015	2,767	1,306	4,604	1,774	957

Notes. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests of Educator Licensure; SPR = summative performance ratings; std = standardized; TOR = teacher of record; VA = value added.

Table 2. Regressions Predicting Novice Teacher Contributions to Summative Performance Ratings

Column	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Summative and Formative Scores									
CAP Summative Score (standardized)	0.077*** (0.022)		0.094*** (0.028)	0.083*** (0.023)		0.076*** (0.026)		0.071*** (0.027)	
CAP Formative Score (standardized)		0.034 (0.022)	-0.025 (0.027)		0.040* (0.022)		0.035 (0.024)		0.021 (0.026)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	1653	1653	1653	1549	1549	1549	1549	1524	1524
R-squared	0.012	0.005	0.012	0.022	0.014	0.083	0.078	0.140	0.134
Panel B: CAP Summative Standard-Level Ratings									
CAP Summative Standard 1	0.073*** (0.021)			0.055* (0.032)	0.054** (0.025)		0.048 (0.032)	0.041 (0.033)	0.035 (0.037)
CAP Summative Standard 2		0.068*** (0.022)		0.023 (0.034)		0.044 (0.028)	0.009 (0.036)	0.025 (0.037)	0.021 (0.040)
CAP Summative Standard 4			0.060*** (0.020)		0.030 (0.025)	0.034 (0.026)	0.028 (0.026)	0.025 (0.026)	0.021 (0.028)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	1653	1653	1653	1653	1653	1653	1653	1549	1524
R-squared	0.011	0.010	0.009	0.011	0.012	0.011	0.012	0.022	0.140
Panel C: CAP Summative Dimension-Level Ratings									
CAP Summative Quality Dimension	0.051** (0.022)			-0.000 (0.032)	-0.014 (0.033)		-0.016 (0.034)	-0.014 (0.035)	-0.015 (0.039)
CAP Summative Scope Dimension		0.072*** (0.022)		0.072** (0.032)		0.006 (0.050)	0.012 (0.050)	0.036 (0.053)	0.057 (0.054)
CAP Summative Consistency Dimension			0.082*** (0.022)		0.092*** (0.033)	0.076 (0.049)	0.083 (0.052)	0.063 (0.054)	0.029 (0.056)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	1653	1653	1653	1653	1653	1653	1653	1549	1524
R-squared	0.007	0.011	0.013	0.011	0.013	0.013	0.013	0.023	0.140

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

Table 3. Regressions Predicting Novice Teacher Value Added

Column	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Summative and Formative Scores									
CAP Summative Score (standardized)	-0.015 (0.013)		-0.021 (0.017)	-0.019 (0.014)		-0.004 (0.014)		-0.008 (0.015)	
CAP Formative Score (standardized)		-0.002 (0.012)	0.009 (0.015)		-0.003 (0.013)		0.008 (0.013)		0.012 (0.014)
MTEL Controls				X	X	X	X	X	X
Provider Fixed Effects						X	X	X	X
Program Fixed Effects								X	X
Observations	891	891	891	848	848	848	848	841	841
R-squared	0.004	0.002	0.005	0.012	0.008	0.145	0.145	0.202	0.203
Panel B: CAP Summative Standard-Level Ratings									
CAP Summative Standard 1	-0.013 (0.013)			-0.005 (0.020)	-0.011 (0.015)		-0.005 (0.020)	-0.005 (0.020)	0.019 (0.020)
CAP Summative Standard 2		-0.014 (0.013)		-0.010 (0.021)		-0.013 (0.017)	-0.010 (0.023)	-0.011 (0.023)	-0.018 (0.024)
CAP Summative Standard 4			-0.008 (0.011)		-0.003 (0.013)	-0.001 (0.014)	-0.001 (0.014)	-0.005 (0.015)	-0.010 (0.015)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	891	891	891	891	891	891	891	848	841
R-squared	0.004	0.004	0.003	0.004	0.004	0.004	0.004	0.012	0.204
Panel C: CAP Summative Dimension-Level Ratings									
CAP Summative Quality Dimension	-0.009 (0.013)			0.003 (0.017)	-0.002 (0.017)		0.002 (0.018)	-0.003 (0.019)	-0.006 (0.019)
CAP Summative Scope Dimension		-0.017 (0.013)		-0.019 (0.017)		-0.024 (0.024)	-0.024 (0.025)	-0.024 (0.029)	0.006 (0.031)
CAP Summative Consistency Dimension			-0.012 (0.013)		-0.010 (0.017)	0.008 (0.023)	0.007 (0.024)	0.007 (0.028)	-0.009 (0.032)
MTEL Controls								X	X
Provider Fixed Effects									X
Program Fixed Effects									X
Observations	891	891	891	891	891	891	891	848	841
R-squared	0.003	0.005	0.003	0.005	0.003	0.005	0.005	0.013	0.203

Notes. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. Outcome is teacher value added to student test score gains, stacked across math and English language arts. *P*-values from two-sided *t*-test: * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

Table 4. Regressions Predicting Novice Teacher Outcomes, Interactions With CAP Year

Column	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Summative Performance Rating Outcome								
CAP Summative Score (standardized)	0.077*** (0.022)	0.074*** (0.028)	0.083*** (0.023)	0.068** (0.028)	0.076*** (0.026)	0.067** (0.030)	0.071*** (0.027)	0.062* (0.033)
CAP Summative Score * CAP Year 2018		0.010 (0.046)		0.052 (0.050)		0.033 (0.051)		0.029 (0.055)
MTEL Controls			X	X	X	X	X	X
Provider Fixed Effects					X	X	X	X
Program Fixed Effects							X	X
Observations	1653	1653	1549	1549	1549	1549	1524	1524
R-squared	0.012	0.012	0.022	0.023	0.083	0.083	0.140	0.140
Panel B: Teacher Value-Added Outcome								
CAP Summative Score (standardized)	-0.015 (0.013)	-0.028* (0.016)	-0.019 (0.014)	-0.029* (0.016)	-0.004 (0.014)	-0.013 (0.015)	-0.008 (0.015)	-0.015 (0.017)
CAP Summative Score * CAP Year 2018		0.051** (0.026)		0.049 (0.032)		0.044 (0.029)		0.032 (0.032)
MTEL Controls			X	X	X	X	X	X
Provider Fixed Effects					X	X	X	X
Program Fixed Effects							X	X
Observations	891	891	848	848	848	848	841	841
R-squared	0.004	0.009	0.012	0.016	0.145	0.148	0.202	0.204

Notes. CAP = Candidate Assessment of Performance; MTEL = Massachusetts Tests for Educator Licensure. *P*-values from two-sided t-test: * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

Table 5. Regressions Predicting All Teacher Outcomes, Interactions With Teacher of Record

Column	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Summative Performance Rating Outcome								
CAP Summative Score (standardized)	0.054*** (0.016)	0.076*** (0.022)	0.060*** (0.017)	0.082*** (0.024)	0.063*** (0.019)	0.075*** (0.025)	0.061*** (0.020)	0.072*** (0.026)
CAP Summative Score * Teacher of Record		-0.047 (0.033)		-0.049 (0.035)		-0.027 (0.037)		-0.027 (0.037)
MTEL Controls			X	X	X	X	X	X
Provider Fixed Effects					X	X	X	X
Program Fixed Effects							X	X
Observations	2742	2742	2476	2476	2476	2476	2456	2456
R-squared	0.008	0.009	0.019	0.020	0.055	0.055	0.113	0.113
Panel B: Teacher Value-Added Outcome								
CAP Summative Score (standardized)	0.001 (0.010)	-0.014 (0.013)	-0.001 (0.011)	-0.019 (0.014)	0.010 (0.011)	-0.005 (0.013)	0.008 (0.012)	-0.009 (0.014)
CAP Summative Score * Teacher of Record		0.031 (0.021)		0.039* (0.022)		0.033 (0.021)		0.037 (0.023)
MTEL Controls			X	X	X	X	X	X
Provider Fixed Effects					X	X	X	X
Program Fixed Effects							X	X
Observations	1287	1287	1173	1173	1173	1173	1159	1159
R-squared	0.001	0.005	0.010	0.015	0.132	0.135	0.179	0.183

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