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

**EXECUTIVE SUMMARY**

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

Center for Analysis of Longitudinal Data in Education Research (CALDER)

American Institutes for Research[[1]](#footnote-1)

Meagan Comb, Elizabeth Losee, Adrienne Murphy, and Aubree Webb

Massachusetts Department of Elementary and Secondary Education (DESE)

Massachusetts introduced the [Candidate Assessment of Performance](http://www.doe.mass.edu/edprep/cap/) (CAP) as a new teacher preparation program completion requirement for teacher candidates during the 2016–17 academic year. CAP is a practice-based assessment of teaching skills aligned with the Massachusetts [Standards for Effective Practice](http://www.doe.mass.edu/edprep/advisories/TeachersGuidelines.pdf) and the state’s associated in-service evaluation rubric. Candidates typically take the assessment under the supervision of an in-service teacher (the supervising practitioner) and a representative from their teacher preparation program (the program supervisor) during the student teaching placement, although about 25% of CAP participants take the assessment as a teacher-of-record to advance certification.

Massachusetts developed CAP to provide a consistent mechanism for evaluating candidate readiness to teach. As a result, CAP has come to embody the state’s increased expectations for preparation providers and novice teachers. CAP also 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 program improvement. For CAP to function as conceived, however, the information that candidates, programs, and the state receive from CAP should predict how candidates will perform in the state’s teaching workforce. DESE, therefore, partnered with CALDER to conduct a research project examining the ability of CAP performance in 2016–17 (the first year of full CAP implementation) to predict in-service performance evaluations in 2017–18. While sample sizes with one year of data precluded an analysis of teachers’ impacts on student achievement (or “value added”), future work will explore the relationship between CAP performance and teacher value added.

Below, the DESE team offers four takeaways related to the results of this research. Within each section, the CALDER research team outlines the primary findings and supporting evidence from this research project, all of which is further described in the full [CALDER working paper](https://caldercenter.org/sites/default/files/CALDER%20WP%20223-1019.pdf). The DESE team then elaborates on the takeaway and research summary, describing additional context and implications for the field of teacher preparation in Massachusetts.

The four main DESE takeaways from this research are:

| [Takeaway 1](#_heading=h.gjdgxs) | Given evidence of CAP’s predictive validity, including evidence that CAP is not differentially predictive for candidates with or without prior teaching experience or from different program types or fields, we will stay the course with CAP as a summative evaluation for initial licensure. |
| --- | --- |
| [Takeaway 2](#_heading=h.30j0zll) | Given that the formative assessment provides additional information about candidate readiness, we will continue to include both summative and formative scores as part of CAP. |
| [Takeaway 3](#_heading=h.1fob9te) | Given a low correlation between CAP and Massachusetts Tests for Educator Licensure (MTEL) performance along with evidence of the predictive validity of each assessment, we will continue to use both CAP and MTEL to provide a combined, comprehensive picture of future teaching effectiveness. |
| [Takeaway 4](#_heading=h.3znysh7) | Given the predictive strength of Scope when all three dimensions are considered simultaneously, we will maintain our current CAP rubrics that evaluate essential elements using more than a single Quality dimension. |

# Policy Takeaway 1: Given evidence of CAP’s predictive validity, including evidence that CAP is not differentially predictive for candidates with or without prior teaching experience or from different program types or fields, we will stay the course with CAP as a summative evaluation for initial licensure.

## Research Summary

*Conclusion: Candidates’ performance on CAP predicts their in-service summative performance evaluations the following year.*

A one standard deviation increase in a candidate’s summative CAP score predicts a 0.11–0.16 standard deviation increase in his or her summative performance rating the following year, controlling for classroom context and other external influences on performance ratings. In comparison, this corresponds to about 40%–60% of the average difference in summative performance ratings between first- and second-year teachers. These relationships are statistically significant and hold whether comparisons are made within or across teacher preparation providers and programs and when models control for candidate scores on the MTEL.

**Figure 1.** CAP Scores and Summative Performance Ratings by Candidate Characteristics

Figure 1. CAP Scores and Summative Performance Ratings by Candidate Characteristics: 
This figure shows the expected change in teachers’ summative ratings associated with a one standard deviation increase in CAP scores, both point estimates and their associated 95% and 80% confidence intervals. All values have been standardized. All candidates, novice teachers, teachers of record, elementary program, SPED program, and Post-Baccalaureate program measures are all significantly positive. While Baccalaureate program candidates’ point estimate is positive, this value is not statistically significant.


*Note.* Dots represent point estimates of the expected change in standardized teacher summative performance ratings associated with a one standard deviation (SD) increase in CAP scores for each group of teachers. The thin line represents the 95% confidence interval, and the thick line represents the 80% confidence interval for each estimate.

*Conclusion: There is no evidence that CAP is differentially predictive for different types of candidates, including, for example, those with or without prior teaching experience or from different program types or fields.*

We test whether CAP is more predictive of future in-service evaluations for candidates with different prior teaching experiences, in different program areas, and in different program types. In each case, we find little evidence that the relationships between CAP and future in-service evaluation scores are stronger for one group than for the other. For example, as shown in Figure 1, the relationships between CAP performance and future performance ratings are similar for novice teachers and teachers of record, for candidates from the two most common types of programs (elementary and special education), and for candidates from Baccalaureate and Post-Baccalaureate programs (although, due to smaller sample sizes, not all individual estimates are statistically significant).

Policy PerspectiveAs Massachusetts seeks to rapidly accelerate and deepen the readiness of candidates for the realities of the classroom, these results indicate that CAP can provide useful and timely feedback about the specific skills and competencies of individual candidates—to the candidates themselves and to their teacher preparation programs. Given this, we plan to stay the course with CAP as a summative evaluation for initial teacher licensure.

The expectations and format of CAP mirror the evaluations used in our PK–12 schools. It is encouraging to see a strong empirical relationship between the two, even after introducing and controlling for various candidate characteristics such as teachers’ fields and levels of prior experience. While data from other states suggest that special education teachers are evaluated differently than their general education peers (Jones et al., 2019)[[2]](#footnote-2), our data does not show a difference in the ability of CAP to predict performance for candidates from different program types (including special education).

CAP scores should not be differentially predictive for teachers with and without prior experience, but the scores should be able to capture that teachers with prior experience are likely to be more prepared for classroom responsibilities than are novice educators. This is borne out in the results of this empirical analysis. Approximately 25% of candidates enrolled in a teacher preparation program were simultaneously employed as full-time classroom teachers and were subject to the PK–12 educator evaluation process in addition to CAP. While the state provides guidance around this facet of CAP implementation, this research provides empirical support that CAP is predictive of future performance for our currently employed candidates as well.

# Policy Takeaway 2: Given that the formative assessment provides additional information about candidate readiness, we will continue to include both summative and formative scores as part of CAP.

## Research Summary

*Conclusion:* *Both the CAP formative and summative performance ratings predict future performance ratings, and the formative assessment relationship is slightly more predictive.*

Teacher candidates receive two ratings during the CAP evaluation cycle. The formative evaluation occurs about halfway through the cycle and is not immediately consequential for program completion. While both the formative and summative CAP ratings predict future in-service performance ratings, this relationship is slightly higher for the formative CAP rating. In other words, if two candidates have the same summative CAP scores, the candidate with higher formative CAP scores is more likely to receive higher performance evaluations in her or his first year of teaching.

## Policy Perspective

This finding about the importance of a candidate’s formative CAP score in predicting later outcomes is novel and important, as it is not something we often have insight into for other large-scale, standardized assessments. While a formative stage is a central component of many evaluation systems nationally, the communication of these formative CAP ratings in real time to candidates—as well as to preparation providers and the state as part of CAP reporting requirements—suggests that these ratings could play an important role in candidate professional development and teacher preparation program improvement in the state. As such, we plan to continue to include formative ratings as part of CAP annual reporting.

It is important for us and our Massachusetts providers to consider the reasons why the formative rating, rather than the summative, is more closely correlated with actual performance once employed. One hypothesis is that CAP evaluators feel pressure during the summative rating stage to signal a level of readiness for candidates that, if stakes were not attached, they may be less inclined to do. We are challenged by this potential scenario given the need to have a summative assessment that is consequential for licensure yet is also an authentic capture of readiness.

Importantly, we want to support our providers to use this information for candidate development and program improvement in order to more actively monitor and assess their candidates before they have classroom responsibilities of their own. One mechanism that we have in place is the Edwin Analytics CAP Dashboard that visualizes CAP ratings for providers using various data displays (see Figure 2 for an example).

# Figure 2. Sample CAP Dashboard Figure 2. Sample CAP Dashboard: The sample CAP Dashboard compares the organization, its programs, and state averages when relevant. The CAP Dashboard includes formative CAP data, a breakdown of each rating in the summative CAP evaluation overall as well as the six elements, a visual display of consistency between programs, and comparisons of CAP to other metrics, i.e. educator evaluation and SGP.

*Note.* Figure displays an example CAP Dashboard for a provider, comparing the organization, its programs, and state averages when relevant. The CAP Dashboard includes formative CAP data, a breakdown of each rating in the summative CAP evaluation overall as well as the six elements, a visual display of consistency between programs, and comparisons of CAP to other metrics.

# Policy Takeaway 3: Given a low correlation between CAP and MTEL performance along with evidence of the predictive validity of each assessment, we will continue to use both CAP and MTEL to provide a combined, comprehensive picture of future teaching effectiveness.

## Research Summary

*Conclusion: CAP provides a signal of future teacher effectiveness beyond what is already captured by existing licensure tests.*

Prior work from other states has shown that basic skills tests and tests of teacher content knowledge, like the different fields of MTEL, provide signals of future teacher effectiveness (e.g., Clotfelter et al., 2007; Goldhaber et al. ,2017).[[3]](#footnote-3) In this analysis, we find that CAP and MTEL scores are only weakly correlated, suggesting that they capture different dimensions of candidate skills. We also find that, when models control for candidate performance on MTEL communication and literacy tests in reading and writing, CAP scores are still significantly predictive of future summative performance ratings. Thus, CAP provides a signal beyond what is being captured by the MTEL scores; i.e., if there are two candidates from the same teacher preparation program with the same MTEL scores, the candidate with higher CAP scores is more likely to receive higher performance evaluations in his or her first year of teaching.

Policy PerspectiveThis finding suggests that, as intended, MTEL and CAP are testing distinct components of candidate readiness. Each provides unique insight into an individual’s knowledge and skill, and each is part of a more comprehensive picture of what is needed to be effective in the classroom. With future updates, we aim to maintain this distinction while adding to the combined, comprehensive picture of future teacher effectiveness. Specifically, we intend to [update CAP](http://www.doe.mass.edu/edprep/cap/2020-21guidelines.docx) to include a new seventh essential element that directly evaluates candidates’ fluency with their respective subject-matter knowledge. This element will build on the expectation of functional content knowledge that is assessed on MTEL and will serve to emphasize the application of subject-matter knowledge and pedagogical practice together. As this element is added and implemented, DESE will continue to monitor the extent to which the CAP and MTEL measurements remain distinct.

# Policy Takeaway 4: Given the predictive strength of Scope when all three dimensions are considered simultaneously, we will maintain our current CAP rubrics that evaluate essential elements using more than a single Quality dimension.

## Research Summary

*Conclusion: CAP performance on different Standards for Effective Practice and rubric dimensions (Quality, Consistency, and Scope) are both individually and as aggregated measures predictive of future summative performance ratings.*

CAP includes essential elements (e.g., Meeting Diverse Needs) from three of the four Standards for Effective Practice used in in-service evaluations. When we group CAP elements into their corresponding standards, we find that they predict teachers’ performance on their future summative performance ratings. For each of these essential elements, evaluators also provide a rating on three dimensions of the candidate’s practice: Quality, Scope, and Consistency. In order to pass CAP, candidates are required to receive a proficient rating on the Quality dimension but can receive a needs improvement rating on the Scope and Consistency dimensions. We find that the dimensions with lower passing requirements are *more* predictive of future performance (Figure 3); and in models that control for scores on all three dimensions at the same time (not shown in Figure 3), these relationships appear to be driven by the Scope dimension. This may reflect the relative importance of this dimension, or it may signify that evaluators are using the less consequential scores to provide additional feedback to candidates.

**Figure 3.** CAP Scores and Summative Performance Ratings by CAP Dimension

**Figure 3. CAP Scores and Summative Performance Ratings by CAP Dimension:
Similar to Figure 1, this figure shows the expected change in teachers’ summative ratings associated with a one standard deviation increase in CAP scores, both point estimates and their associated 95% and 80% confidence intervals. All values have been standardized. Estimates are significantly positive for all dimensions: Overall, Quality, Scope, and Consistency.
**

*Note.* Dots represent point estimates of the expected change in standardized teacher summative performance ratings associated with a one standard deviation (SD) increase in each CAP score. The thin line represents the 95% confidence interval, and the thick line represents the 80% confidence interval for each estimate.

*Conclusion: CAP evaluators are not using the full range of evaluation scores available to rate candidates.*

CAP provides opportunities for feedback in six standards and three dimensions for each standard, for a total of 18 ratings. About one third of all candidates receive a proficient rating in every single category (represented by the spike at a summative score of 54 in Figure 4), and almost no candidates receive an unsatisfactory rating on any category.

**Figure 4.** Distribution of Summative CAP Scores

**Figure 4. Distribution of Summative CAP Scores:
The distribution of summative CAP scores ranges from 42 to 72 with a sharp spike at 54, the score if all measures are scored as “proficient”.
**

## Policy Perspective

The dimensions of Quality, Scope and Consistency are underlying constructs for performance descriptors in the state’s model [rubrics](http://www.doe.mass.edu/edeval/resources/rubrics/). In designing CAP, we intentionally pulled these constructs of practice apart in an effort to differentiate the various ways in which a candidate may be fully meeting expectations while also still developing her or his skills as a novice teacher. Some providers have noted that the differences between these dimensions are both not necessary (i.e., creating more work) and difficult to distinguish in practice (i.e., creating challenges with inter-rater calibration). The research findings above suggest to us that the dimensions individually provide meaningful information about candidate performance, and the ratings associated with the constructs with lower passing requirements are perhaps even more informative for candidates. Thus, we intend to maintain the current breakdown by dimension, as simply looking at the Quality dimension is likely insufficient to assess readiness alone, and it appears that candidates receive valuable feedback from evaluators on the Scope and Consistency scores.

Looking at Figure 4, we are interested in further examining the implications of thresholds on rating practice. Removing the Scope and Consistency dimensions or changing all three dimensions to Proficient thresholds would likely result in even greater consolidation, limiting the usefulness of the scores. Evidence suggests that evaluators may use the less consequential Scope and Consistency dimensions to provide feedback to teachers about their progress, but if they were doing so more regularly, we would see a spike at a score of 42 rather than 54. We plan to make no changes to the current passing requirements and encourage CAP evaluators to use the full range of scores available to them. To this end, we will continue to encourage preparation providers to use the [Online Platform for Teaching and information Calibration (OPTIC)](http://www.ma-optic.com/) to guide and assess the quality and consistency of judgements made by supervisors.

Combining this with the other research findings in this paper, we recognize the value of both summative and formative CAP scores along three dimensions and will continue to support providers in their efforts to give timely, calibrated feedback to candidates about their progress—and to use this data to inform preparation program improvement as well as our own state policy.

# Areas for Further Inquiry

In the next year of our partnership with CALDER, we will explore similar questions about CAP through the lens of student achievement results. The aim is to examine the relation between educators’ CAP performance and their impact on student achievement. We will continue to monitor our data and research results and to revisit the policy and implementation practices associated with CAP as necessary.

1. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H170025 to the American Institutes for Research (AIR). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)
3. 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.

   Goldhaber, D., Gratz, T., & Theobald, R. (2017). 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. [↑](#footnote-ref-3)