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

Teacher Licensure Tests: Barrier or Predictive Tool?

James Cowan
Dan Goldhaber
Zeyu Jin
Roddy Theobald

WORKING PAPER No. 245-1020 • October 2020

Teacher Licensure Tests: Barrier or Predictive Tool?

James Cowan
American Institutes for Research/CALDER

Dan Goldhaber
American Institutes for Research/CALDER

Zeyu Jin
American Institutes for Research/CALDER

Roddy Theobald
American Institutes for Research/CALDER

Contents

Contents i
Acknowledgments ii
Abstract iii
1. Introduction 1
2. Background Literature and the Massachusetts Context 3
2.1 Licensure Tests and Teacher and Student Outcomes 3
2.2 Massachusetts Context 5
3. Data and Measures 6
3.1 Candidate and Teacher Data 6
3.2 Candidate Race/Ethnicity Data 6
3.3 Student Achievement Data 7
3.4 Teacher Evaluation Data 8
4. Analytic Approach 9
5. Results 10
5.1 Main Results 11
5.2 Nonrandom Student-Teacher Sorting 13
5.3 Sample Selection Bias 14
6. Conclusion 16
References 19
Tables and Figures 24
Appendix A 33

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H170025 to the American Institutes for Research (AIR). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. This research was also supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. The authors wish to thank Bingjie Chen and Sydney Payne for outstanding research assistance. We also thank Meagan Comb, Michael Hansen, Liz Losee, and Aubree Webb, for comments on an earlier draft of this paper.

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

Teacher Licensure Tests: Barrier or Predictive Tool?
James Cowan, Dan Goldhaber, Zeyu Jin, Roddy Theobald
CALDER Working Paper No. 245-1020
October 2020

Abstract

We use teacher candidate test scores on the Massachusetts Tests for Educator Licensure (MTEL), linked to student and teacher outcomes in the state, to investigate the predictive validity of these teacher licensure tests. We find that MTEL scores are positive and statistically significant predictors of teachers' in-service performance ratings and contributions to student test scores (i.e., value added) once they enter the workforce. We then explore whether these relationships vary for candidates and teachers of color. We find that teacher candidates of color have lower first-time pass rates and are also less likely to retake licensure tests if they fail than are White teacher candidates, but we do not find consistent evidence that MTEL scores are less predictive of value added for teachers of color. Finally, we find that MTEL scores are more predictive of teacher performance ratings for teachers of color than for White teachers.

1. Introduction

Licensure testing of prospective teachers is one of the primary tools that states use to ensure that teachers have a basic level of competence. Testing as a condition for teacher employment eligibility is both ubiquitous and longstanding. Licensure testing began in Pennsylvania in 1834 (Ravitch, 2003), and has been regularly used by most states since the 1960s (Angrist and Guryan, 2008). In 2020, all states require prospective teachers to pass one or more licensure tests to be eligible to teach.[1]

The credentialing role of testing requirements has become increasingly controversial, however. Tests are a potentially costly barrier to entering the teaching profession and only modestly correlated with in-service teacher performance measures (Ballou & Podgursky, 1998; Gershenson et al., 2021; Petchauer, 2019). Black and Hispanic teacher candidates are also substantially less likely to pass these tests (e.g. Goldhaber et al., 2017; Rucinski & Goodman, 2019).[2] Licensure tests and cut scores differ by state, so there is no comprehensive national picture, but evidence from commonly used tests shows that the gap in first time passing rates is 30-40 percentage points between Black and White teacher candidates and 15-25 percentage points between Hispanic and White teacher candidates, depending on the test type (Nettles et al., 2011; Tyler et al., 2011).[3]

This disparate impact of testing, combined with increased evidence that there are academic benefits to students of color with racial and ethnic teacher role models (Dee, 2004, 2005; Egalite et al., 2015; Gershenson et al., 2018),[4] has raised concerns that testing requirements limit the diversity of the teacher workforce, perhaps with little effect on the competence of new teachers. Numerous stories about the disparate impact of testing have appeared in national media outlets, such as the *New York Times* (Harris, 2015), and there have also been a number of high-profile lawsuits questioning whether licensure tests are discriminatory.[5]

There has also been recent legislative action on licensure tests. Washington State, for instance, eliminated basic skills cut score requirements in 2019 (prospective teachers are still required to

[1] In some states, under alternative licensure provisions, passing licensure tests represents the sole criterion, other than a college degree and background check, that individuals must satisfy to be eligible to teach.

[2] There are a number of potential explanations for these differences in performance, from the "Eurocentric" focus of licensure tests (Sleeter, 2017) to opportunity gaps between White candidates and candidates of color in terms of prior educational experiences (Nettles et al., 2011).

[3] Licensure tests are also controversial given that they were historically used by some states and school districts to rationalize racial inequities in teacher salaries and displace Black teachers in the wake of desegregation and equalization orders (Baker, 1995; Tillman, 2004; Smith, 1988).

[4] For more context about the history of testing and the potential tradeoffs between licensure testing and the diversity of the teacher labor market, see (Gershenson et al., 2021).

[5] Historically, courts have come down on different sides of lawsuits challenging testing requirements. For instance, courts have turned away challenges of teacher licensure procedures in California, Alabama, and South Carolina (Mitchell et al., 2001). By contrast, a 2012 ruling in *Gulino v. Board of Education*, for instance, found that a licensure test used in New York City violated Title VII of the Civil Rights Act because the test used by the district had a disparate racial impact and was not demonstrated to be related to relevant job requirements (Center for Constitutional Rights, 2012).

take the state's basic skills tests and to pass a test of subject matter knowledge), and California is on the precipice of eliminating a test on reading instruction (Lambert, 2020). The widespread suspension of licensure tests due to the COVID-19 pandemic is likely to heighten debates about the efficacy of licensure tests because there will be many more teachers in the workforce who have not taken or passed these tests (Skinner et al., 2020).

The efficacy of licensure testing requirements has enormous public policy implications. There are approximately 100,000 new teachers each year, most of whom would be required to take licensure tests (Cowan et al., 2017), therefore tests can have a major impact on the employment eligibility of a significant number of prospective public sector workers. There is also a large discrepancy between the demographics of teachers and students in public schools: As of 2016, about half of public school students were non-White, compared to only 20% of teachers (USDOE, 2017). This *diversity gap* has been identified as a potential source of academic disparities among K–12 students (Partelow et al., 2017; Weir, 2016).

Determining the overall effects of licensure testing requirements is difficult, given that these requirements could impact who pursues teaching as a career, how teacher candidates prepare, and which teacher applicants are offered positions and employed in public schools (Angrist & Guryan, 2008; Boyd et al., 2007; Goldhaber, 2004). That said, licensure testing requirements are certainly more likely to have a positive impact on the quality of the teacher workforce if they are strongly predictive of teacher effectiveness (Goldhaber, 2011). Moreover, judgements about the disparate impact of licensure tests depend in part on the external validity of these tests for White teacher candidates and candidates of color.

We investigate the predictive validity of licensure tests—the extent to which test scores predict in-service performance evaluations and contributions to student achievement—using linked teacher and student administrative data from Massachusetts and licensure test scores on the Massachusetts Tests for Educator Licensure (MTEL). We find that candidates' performance across different MTEL fields is a positive and statistically significant predictor of their in-service performance ratings and contributions to student test scores (i.e., value added) once they enter the workforce.

We also test whether these relationships vary for teachers of color. We find significant differences by teacher race/ethnicity in licensure test scores, with teacher candidates of color having lower first-time pass rates. Candidates of color are also less likely to retake licensure tests if they fail, and their decisions about retakes are more sensitive to their initial score than are the decisions of White candidates. Although teachers of color score lower on licensure tests, we do not find consistent evidence that tests are less predictive of the impacts that teachers of color have on student tests in math or English language arts (their value added) than White teachers. This finding runs counter to prior findings on the Praxis in North Carolina (Goldhaber & Hansen, 2010), and, importantly, the precision of our estimates allows us to rule out even relatively small differences in the predictive validity of the MTEL by teacher race/ethnicity.

Finally, we assess the differential predictive validity of licensure tests for teacher performance ratings, which has, to our knowledge, not previously been assessed. This is important, given that performance ratings are available for a much larger sample of teachers and may better capture the totality of teacher contributions to students.[6] Here, we find that licensure tests are *more* predictive of teacher performance ratings for teachers of color than for White teachers in Massachusetts. These findings are robust to multiple tests of bias that could arise from nonrandom student sorting or sample selection.

2. Background Literature and the Massachusetts Context

2.1 Licensure Tests and Teacher and Student Outcomes

Few studies attempt to estimate the causal effects of licensure testing requirements on the composition of the teacher workforce. The existing studies rely on difference-in-difference designs using a major expansion in state testing requirements beginning in the early 1980s. Angrist and Guryan (2004, 2008) use data from the Schools and Staffing Survey to assess how the introduction of state licensure testing requirements affected the qualifications of new teachers. They find that licensure testing requirements increased teacher wages but had little to no effect on the observable qualifications of teachers. Notably, they did find testing requirements reduced the number of Hispanic (but not Black) teachers. However, it is somewhat uncertain whether these studies identify the effects of certification laws, given the lack of large national data sets on teacher qualifications and student achievement and the possibility that states implement other school reforms alongside changes in licensure rules.[7]

A larger number of descriptive studies use statewide longitudinal data and find that teacher performance on licensure tests predicts student test achievement (Clotfelter et al., 2007, 2010; Goldhaber, 2007; Goldhaber & Hansen, 2010; Goldhaber et al., 2017; Hendricks, 2015 Sass, 2015). For example, Clotfelter et al. (2007, 2010) find that a one standard deviation increase in licensure test scores is associated with improvements in student achievement of about 0.005–0.015 standard deviations.

Importantly, however, there is some evidence that the predictive validity of tests varies for different students; for example, Vars and Bowen (1998) find that SAT scores are less predictive of college outcomes for Black students than for White students. The one prior study that has investigated this question for teacher licensure tests, Goldhaber and Hansen (2010), finds some evidence that licensure test scores are differentially predictive of the future effectiveness of White and Black teacher candidates in North Carolina. They find that a test of pedagogy and content knowledge appears to be more predictive of future performance for White teachers and that a test of content knowledge alone appears to be more predictive of performance for Black

[6] See, for instance, Backes and Hansen (2018), Jackson (2018), and Kraft (2019), for research on teachers' contributions to non-test student outcomes.

[7] For instance, Massachusetts introduced the licensure tests studied in this paper following comprehensive school reform legislation targeting school finance, curriculum, and assessment.

teachers.

The potential that licensure tests are differentially predictive, or perhaps not predictive, of future effectiveness of some types of teachers, is an important policy issue, given the substantial evidence that licensure testing requirements have disparate impacts on eligibility to teach (Goldhaber & Hansen 2010; Rucinski & Goodman, 2019). In Massachusetts, Rucinski and Goodman (2019) find that the teacher pipeline is less diverse than the student population at all stages and that differences in first-time pass rates on the MTEL Communication and Literacy Skills tests contribute to but are not the most important factors in the lack of diversity in the candidate pool.

In this study, we focus on biases in the predictive validity of licensure tests for predicting in-service performance. Because licensure tests seek to identify teachers who are low performing, understanding whether the predictive validity of the tests differs across groups of teacher candidates is clearly important. But it is important to note that licensure tests may additionally suffer from several other kinds of biases.

Licensure tests may suffer from several forms of methodological or content biases that make them less favorable to candidates of color (Jencks, 1998). For instance, the tests might cover segments of the state curricular framework or rely on assessment formats that are more difficult for candidates from some cultural backgrounds. In addition, the phrasing of individual questions may affect performance for candidates of color (Dee & Domingue, 2019).[8] If this were the case, then we might expect lower pass rates for teachers of color even if they do not differ in true content knowledge. These forms of bias may or may not manifest as differences in the predictive validity of the tests. Although we do not address them in this study, the existing MTEL development protocol does include tests of racial bias on individual items. We describe the bias review process in more detail in the next section.

In addition to concerns about biases in the selection and construction of individual items, the scope of licensure testing requirements may also lead to disparities in passing rates. By design, licensure tests cover a subset of the skills that matter for teaching, and these teaching skills may disproportionately favor candidates from certain backgrounds. For instance, several national licensure exams, such as the Praxis test considered by Goldhaber and Hansen (2010), test pedagogy in addition to content knowledge. Candidates attending traditional preparation programs as part of their undergraduate education may have more understanding of pedagogical theory than candidates with more subject-specific experience who enter the profession through alternative pathways. Given that there are significant differences in race across licensure pathways (Cowan et al., 2018), some tests might privilege White teachers over equally effective teachers of color. As is the case with content biases, the tests of predictive validity we conduct in

[8] Dee and Domingue (2019) study a question on the Massachusetts Comprehensive Assessment System (MCAS) in spring 2019 that asked students to write an essay from the perspective of a character with racist beliefs. They found that Black students performed worse on *subsequent* questions on the MCAS, although they also present some evidence that their method for detecting bias over-rejects the null hypothesis of no testing bias.

this study may not address these forms of selection system bias (Jencks, 1998). In particular, if teachers of color differ in other skills not well captured by licensure tests, then testing requirements may disproportionately exclude potentially effective teachers.

2.2 Massachusetts Context

In Massachusetts, the setting of this study, the state requires applicants for a prekindergarten (PK)–12 educator license to pass the Massachusetts Tests for Educator Licensure (MTEL) in Communication and Literacy Skills (Reading and Writing) and in at least one additional PK–12 academic subject area when available. The MTEL are aligned to state regulations governing the subject matter knowledge expectations for teacher candidates and state curriculum frameworks. In their current design, most MTEL tests contain a set of subareas that correspond to broad academic subject areas.[9] For instance, the General Curriculum–Multi-Subject test includes subareas for Language Arts, History and Social Science, Science and Technology/Engineering, and Integration of Knowledge and Understanding. The subareas form the general structure of the assessment and determine how heavily each content area is weighted in the calculation of scores. Each subarea contains several objectives that correspond to finer academic areas or tasks and guide the construction of individual test items. These objectives have been approved by technical review panels following a review of state regulations, curriculum frameworks, and surveys of teacher candidates and university faculty. The content advisory committees assign weights to each objective or subarea, which determine the number of questions as well as the aggregation of scores.

Following the development of test items, the advisory panel determines cut scores for each section on each test using an Angoff method. At qualifying score conferences, panelists are first asked to use their judgment to determine what percentage of hypothetical "just acceptably qualified candidates" would answer each question correctly. The panelists are then given information about candidate performance on a pilot exam and allowed to revise their estimates. The test vendor takes the median of these estimates across panelists to calculate the "performance levels" of each item, and these performance levels are then summed to the section level to obtain a panel-recommended cut score for a given section and test. Finally, the vendor incorporates this information into a recommendation to the commissioner, who makes a final determination on the passing scores.

The state also convenes a bias review committee to assist the advisory committee in assessing testing items for potential bias. The bias review committee reviews test content and language for topics or wording that might disadvantage certain populations of teacher candidates. The review committee also checks the test items for offensive or stereotyped content. The bias review committee submits any concerns to the advisory committee, which is required to address all issues raised during the bias review.

[9] The two Communication and Literacy Skills tests required of all candidates for PK–12 licensure in Massachusetts are an exception to this structure, and each test includes only a single subarea.

The bias review process is intended to address some forms of bias described in Section 2.1. The review of content and test language is designed to address the kinds of labeling or content bias that might disadvantage teachers of color on the MTEL. But the review process does not address the external validity of the MTEL nor bias that might arise from testing knowledge of curricular frameworks as a requirement of teacher licensure. This study contributes to the evidence on disparities in licensure testing by assessing whether the external validity of the test varies by teacher candidates' self-reported racial identity.

3. Data and Measures

3.1 Candidate and Teacher Data

Data on teacher candidate MTEL scores come from teacher licensure data provided by the Massachusetts Department of Elementary and Secondary Education. These data include a complete history of candidate scores across all 36 MTEL fields dating back to 1998. We standardize these scores by field and year across all test takers. These data include scores on all test attempts for each candidate. The first-time pass rates vary from about 40% (political science) to over 90% (for some of the foreign language tests).

Our primary specifications use candidates' average scores across all MTEL fields they have taken, but we also consider alternative specifications that consider candidate performance just on the Communication and Literacy Skills fields that are required of all candidates for teacher licensure. Because decisions about whether to retake the test may be related to the potential effectiveness of teacher candidates, in both cases we use the first (rather than highest) score in our preferred analyses. However, the results are similar using highest scores instead (see Appendix Table A.2).

The MTEL data also include information about candidate race/ethnicity. Candidates are asked to identify as American Indian or Alaskan Native, Asian or Pacific Islander, Black, Hispanic, White, or Other. For teachers reporting different identities in separate test administrations, we use the modal category. About 5% of candidates do not report a racial identity on any test administration. For most analyses, we code all teachers who report a racial identity other than White as teachers of color.

We connect these MTEL scores to outcomes observed for teachers in the state's Education Personnel Information Management System (EPIMS). EPIMS includes information about teacher assignments, district evaluation data, and licensure and education status, and it is our main source of information on teacher employment data. We consider outcomes for candidates who are eventually reported in EPIMS as working in a teaching position in a public school in Massachusetts. In Section 5.3, we discuss methods for addressing the sample selection bias that might arise from considering employed teachers.

3.2 Candidate Race/Ethnicity Data

The primary analysis uses self-reported race/ethnicity from the MTEL. We also observe district-

reported race/ethnicity in the EPIMS data for teachers employed in Massachusetts public schools. The MTEL data have the advantage that the race/ethnicity question is asked of all candidates in a standardized format. According to state officials, methods for collecting data on employee race/ethnicity in EPIMS likely vary across school districts and may not be reported by teachers themselves (Weller & Marino, 2020, personal communication).

We identify significant differences between the racial identities reported in the MTEL and EPIMS data sets (**Appendix Table A.1**). Among candidates in both data sets, teachers are more likely to identify as teachers of color in the testing records (12%) than in the administrative data (10%). Of the teachers of color identified in the testing data, 22% are identified as White in the administrative data. Differences in the question format may partially explain these discrepancies. The administrative data set uses a separate Hispanic origin question and the five race groups included in the Office of Management and Budget (1997) guidelines (American Indian or Alaskan Native, Asian, Black, Native Hawaiian or Pacific Islander, White). Research by the Census Bureau suggests that including Hispanic origins in a single combined race/ethnicity question and including an "other race" option (as on the testing form) decreases the number of respondents identifying as non-Hispanic White (Compton et al., 2013). We find some evidence consistent with this possibility. The changes in identified race are most significant for candidates identifying as Hispanic (21%) or as other race (58%) in the MTEL data. However, if the MTEL data include teachers incorrectly identifying as teachers of color, measurement error in the race/ethnicity data will tend to attenuate estimates of differential predictive validity between groups. In Section 5.4, we show that the results are not sensitive to the use of EPIMS race/ethnicity data.

We report differences in MTEL performance (according to the first-time test score measures) by the race/ethnicity reported in the MTEL data in **Table 1**. Hispanic candidates score about 0.63 standard deviations lower than the mean across both CLSTs, and Black candidates score about 0.68 standard deviations lower than White candidates, while the percentage of candidates who pass both CLSTs on the first attempt is about 77% for White candidates, about 54% for Hispanic candidates, and about 50% for Black candidates. The differences in performance are similar on the other subject-specific MTEL fields (Panel B).

3.3 Student Achievement Data

We use data on student achievement in math and English language arts (ELA) for students in Grades 4 through 8 between 2012 and 2019. We link students and teachers through common course codes for class sections. We combine these linked data with test results on the MCAS and Partnership for Assessment of Readiness for College and Careers (PARCC) annual end-of-grade tests. Some schools administered the PARCC assessments in 2015 and 2016. Given evidence of significant online test mode effects on the PARCC assessments (Backes & Cowan, 2019), we include controls for test mode in our models. We standardize test scores by grade and year using a normal curve equivalent transformation, given that the MCAS is scaled using a nonlinear transformation of the estimated student true scores in some years (Jacob & Rothstein, 2016).

We present summary statistics for the sample of teachers for whom we can estimate teachers' value-added impacts on student test achievement, the "value added samples," in **Tables 2** and **3**. In the math sample, we observe 13,957 teachers with MTEL scores (11,671 White teachers, 1,264 teachers of color, and 1,022 teachers with missing race/ethnicity data). Teachers of color tend to be assigned to more students who are disadvantaged than White teachers in this sample. For example, 57% of students assigned to a teacher of color qualify for subsidized lunches, compared to 35% for White teachers. Similarly, baseline achievement is lower for students assigned to teachers of color by about 0.22 standard deviations. The sample sizes and patterns of observable characteristics are similar for ELA teachers (Table 3).

3.4 Teacher Evaluation Data

We use performance evaluation data collected under the Massachusetts state evaluation framework as an additional measure of teacher quality. The state evaluation framework covers four professional teaching standards, and districts evaluate teachers on each of the standards and then create a final summative performance measurement based on their professional judgment of the teacher's entire practice. Although the final summative ratings are associated with student achievement gains (Cowan et al., 2020), there is somewhat limited variation in teacher performance under the summative performance measurement: About 85% of teachers receive a *proficient* rating in this system (Cowan et al., 2020). Fortunately, there is considerably more variation in the individual standard ratings because these ratings do not directly contribute to the final rating. We follow Kraft et al. (2020) and fit a graded response model to the four professional standards ratings, which permits the difficulty and discrimination of each standard to differ. Specifically, for standard *j* and rating level *k*, we estimate

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

In Equation (1), 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 resulting empirical Bayes estimates of θ_i as our measures of teacher performance ratings.

We merge the teacher performance data to samples of matched classrooms in the 2014–19 school years using the linked schedule data. We retain courses that are identified as core subject courses in math, ELA, social studies, and science in Grades K–12. We then collapse the data to the classroom level so that each observation is a single classroom assignment. We present summary statistics for this sample in Table 4. We observe 53,613 teachers with MTEL scores in this sample (43,345 White teachers, 5,337 teachers of color, and 4,931 teachers with missing race/ethnicity data). The general patterns of observable student characteristics are similar as the value-added samples, with teachers of color having more students who are economically disadvantaged, Black and Hispanic, and lower achieving.

4. Analytic Approach

We use standard value-added approaches to estimate the relationship between MTEL scores and student achievement (e.g., Clotfelter et al., 2007; Goldhaber et al., 2017) or teacher performance ratings. We first construct estimates of teacher quality by regressing student achievement or performance ratings Y_{ijt} on student controls X_{ijt}:

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

In Equation (2), the control vector X_{ijt} includes a cubic polynomial in lagged test scores in mathematics 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 PARCC and PARCC online assessments.[10] In models involving teacher evaluations, we additionally include an indicator for a formative assessment and interact grade fixed effects with course subject.

Experimental and quasi-experimental evidence suggests that value-added estimates, properly specified, produce estimates of teacher contributions to student learning with limited bias (i.e., estimates that comport to findings when teachers are randomly assigned to classrooms within schools [Bacher-Hicks et al., 2019; Kane & Staiger, 2008; Kane et al., 2013] and when individual teachers move across schools and grades [Chetty et al., 2014]). Cowan et al. (2020) additionally find that similar methods for adjusting teacher performance ratings produce measures of teacher quality that are nearly forecast unbiased.

Nonetheless, we test the robustness of our results to several additional specifications. We control for two lags of prior test scores, which Rothstein (2009) suggests reduces bias in value-added estimates. We also use multiple methods to adjustment for school effects on student achievement. In some specifications, we include a school-by-grade effect in Equation (2), which implicitly compares a teacher's performance to others in the same school and grade. Because there are often few teachers in a school-grade cell, we also take an intermediate approach to adjusting for school effectiveness using a grouped fixed-effects method (Bonhomme & Manresa, 2015; Bonhomme et al., 2019). We first estimate a naïve version of Equation (2) and average residuals at the school-grade level. We then stratify the sample into groups based on the estimated school value added and include group fixed effects in Equation (2) when constructing teacher value-added estimates. In practice, this means we replace indicators for each school-grade cell in our sample with three indicators (identifying four groups) of schools with similar observed achievement gains. This approach thus compares teachers to other teachers in schools and grades with similar value added to student test scores. Results using the grouped fixed

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

effects and school fixed effects are generally similar.[11]

We then use residuals from Equation (2) to estimate variants of the following analytic model:

$$\epsilon_{ijt} = MTEL_j\beta + \nu_{ijt} \qquad (3)$$

In Equation (3), ϵ_{ijt} is a residual from the value-added model in Equation (2), and $MTEL_j$ is an MTEL score for teacher *j*; in our primary specifications, this is the average score across all MTEL fields. We weight all regressions by the inverse number of observations so that estimation is numerically equivalent to using estimated teacher value added from Equation (2) as the dependent variable. The coefficient of interest, β, represents the expected increase in the measure of teacher effectiveness associated with a one standard deviation increase in the MTEL score. We cluster standard errors at the teacher level to account for dependence across multiple observations for the same teacher.

In addition to assessing the predictive validity of the MTEL overall, we also assess whether its accuracy in predicting workforce outcomes varies for teachers of color in Massachusetts. We assess differential predictive validity by adding an indicator for candidates of color, C_j, and an interaction of this indicator with MTEL score to Equation (3) above:

$$\epsilon_{ijt} = C_j\alpha + MTEL_j\beta + MTEL_j * C_j\gamma + \nu_{ijt} \qquad (4)$$

The coefficient γ can be interpreted as the difference in the relationship between MTEL and our effectiveness measures between White candidates and candidates of color. If γ is positive, then the MTEL is more aligned with teacher effectiveness measures for candidates of color than for White candidates. Similarly, a negative coefficient indicates a weaker relationship.

5. Results

We discuss the main results about the relationship between teachers' licensure test (MTEL) performance and their in-service performance in the following section and then assess several potential sources of bias, including: nonrandom sorting of students to teachers that may differ by teacher race/ethnicity (Section 5.2); and nonrandom entrance into and attrition from the analytical samples (Section 5.3). We generally find that results are not sensitive to modeling decisions.

Before discussing these results, we present some relationships from the first-stage models (Equation 2) that help contextualize the magnitude of the relationships discussed in the rest of this section. All else equal, students eligible for free or reduced priced lunch (FRPL) score about .06 standard deviations lower in math and ELA than students not eligible for FRPL, while

[11] We use the recommendation for determining the number of groups in Bonhomme et al. (2019). We estimate the variance of school value-added estimates accounting for classroom and annual performance shocks. The optimal number of groups is four in each sample, although estimates are not sensitive to using more or fewer groups. We stratify the sample using a k-means clustering algorithm, which minimizes the within-group variance in school value added.

students with a novice teacher (i.e., with no prior teaching experience) score about .04 standard deviations lower in math and about .03 standard deviations lower in ELA than students with a second-year teacher. And, all else equal, a second-year teacher tends to score about .12 standard deviations higher on summative performance ratings than a first-year teacher.

5.1 Main Results

We present the main results on teacher performance and licensure test scores in **Table 5**. Using data on all teachers in our sample (columns 1–3), we estimate statistically significant and positive relationships between average first-time MTEL scores and each of the performance measures included in the study. In column 1, we present results from models that adjust only for student/classroom observables. We add the grouped fixed effects to the value-added specification in column 2 and then replace these grouped fixed effects with school-by-grade fixed effects in column 3.

The results on teacher value added are quite similar to results from prior research (e.g., Clotfelter et al., 2007; Goldhaber, 2007). For instance, as we show in Panel A, a one standard deviation in MTEL performance is associated with an improvement in student test scores of about 0.024 when adjusting only for observable characteristics of students; the magnitude of this coefficient is less than half of the regression-adjusted gap in test performance between FRPL and non-FRPL students and more than half of the expected returns to the first year of teaching in terms of student test performance. This relationship is somewhat weaker (0.012–0.015) when we include group or school-grade effects but is still statistically significant.[12]

As has been found in the previously cited research, the relationship between licensure test scores and ELA achievement is somewhat weaker (Panel B). In models that include grouped or school-grade effects, we estimate that a one standard deviation increase in average MTEL scores is associated with an increase of about 0.005–0.008 student standard deviations in test performance.

The findings for teacher performance evaluations are reported in Panel C. The results are similar across specifications and show that a one standard deviation increase in average MTEL scores corresponds to higher performance ratings of about 0.08–0.10 standard deviations. The magnitude of this relationship is therefore similar to that between math value added and MTEL scores, and it is also more than two thirds of the expected improvement in performance ratings between a teacher's first and second years of teaching.[13]

As we show in **Figure 1**, the relationships between MTEL scores and the in-service performance

[12] By point of comparison, we estimate that one standard deviation change in teacher value added to math and ELA achievement is about 0.15 student standard deviations.

[13] As another point of comparison for these results, Chen et al. (2019) consider a different *preservice* test in Massachusetts, the Candidate Assessment of Performance (CAP) and find the relationships between CAP scores and summative performance ratings are generally stronger than the corresponding relationships for MTEL. This is not terribly surprising, given that the CAP is explicitly designed to mimic the evaluation process that results in the summative performance ratings that in-service teachers receive.

measures (both value-added and summative performance ratings) are approximately linear. We find little evidence of nonlinearities in the licensure scores near the tails of the distribution. However, we note that we observe relatively few teachers who are low performing in the employed sample. We return to this empirical issue in Section 5.3.

In the remaining columns of **Table 5**, we focus on the potential of differential predictive validity between White teachers and teachers of color. We focus mainly on the results with grouped or school-grade effects (columns 5 and 6), given that these effects seem to explain some of the relationships between MTEL scores and the in-service performance measures in columns 1–3. In contrast to Goldhaber and Hansen (2010), who find that licensure test performance is more predictive of student test achievement for White teachers than Black teachers, we find little evidence of differential predictive validity of MTEL scores for math achievement (Panel A).[14] The point estimates on the interaction between teachers of color and MTEL scores (<0.001–0.002) are small and statistically insignificant. For ELA, the estimated interactions between MTEL and teachers of color are somewhat larger and negative but not close to statistically significant. An important caveat is that there are relatively few teachers of color in the sample and the interaction terms are not precisely estimated. Thus, given the estimated confidence intervals, we cannot rule out the possibility that MTEL scores are not correlated with ELA achievement among teachers of color (as shown by the relatively flat regression line for teachers of color in Figure 2).

In Panel C, we assess differential predictive validity for teacher performance ratings. We find that MTEL scores are more predictive of in-service performance for teachers of color, although the interaction terms are only significant at the 10% level. Taken at face value, the point estimates suggest that the relationship between MTEL scores and performance ratings is about 25% larger for teachers of color. As in Figures 1 and 2, we plot the performance measures against MTEL scores separately by teacher race/ethnicity in **Figure 3** and find little evidence of nonlinear relationships.

It is not clear why the relationship for teacher performance ratings, which rely on human judgment, differ from the relationships for standardized test scores. One possibility is that, as shown by Cowan et al. (2020), the variation in performance ratings differs across schools and districts in Massachusetts. If teachers of color tend to work in schools that provide more variable performance ratings, then we would expect to see a stronger relationship between proxies for teacher effectiveness (such as the MTEL) and performance evaluations. When we explore this possibility, we find that the within-school standard deviation of performance ratings is about 14% larger for teachers of color than for white teachers, which would be enough to explain about half of the observed difference in predictive validity. Another possibility is that the performance evaluations suffer from discriminatory biases (e.g., Drake et al., 2019; Grissom & Loeb, 2017).

[14] As shown in Appendix Table A.6, we similarly find no significant differences in predictive validity between White and Asian/Pacific Islander, Black, or Hispanic teachers. The comparison between White and Black teachers in Appendix Table A.6 is analogous to the comparisons made in Goldhaber and Hansen (2010).

If low-performing teachers of color are more likely to receive low performance evaluations than low-performing White teachers, then we might observe a stronger relationship between MTEL and performance evaluations for teachers of color. As in Figures 1 and 2, we plot the performance measures against MTEL scores separately by teacher race/ethnicity in **Figure 3** and find little evidence of nonlinear relationships.

In **Appendix Table A.3**, we conduct similar analyses using the two MTEL tests required of every teacher (Communication and Literacy Skills tests in Reading [Panel A] and Writing [Panel B]), as well as an additional set of models that are pooled across the other subjects tests required for specific licenses (Panel C). The relationships between MTEL scores and teacher performance are generally similar, both in magnitude and statistical significance, to results using the MTEL tests averaged across all the various fields taken by each teacher candidate. The interactions between teachers of color and MTEL scores are only statistically significant and positive for the MTEL subject tests predicting SPR.

5.2 Nonrandom Student–Teacher Sorting

One concern about the models described in Section 5.1 is that the relationships between student *unobservables* and MTEL scores could differ by teacher race/ethnicity. We showed in Tables 2–4 that teachers of color are more likely to teach students who are lower achieving and less advantaged. If there are similar differences in unobserved determinants of student test scores that are differentially correlated with MTEL scores by teacher race, then we may not consistently estimate the relationship between licensure test scores and teacher performance in the classroom. It is not obvious which direction this sort of sorting bias might take. On the one hand, if teachers of color who score higher on the MTEL are systematically matched to students from disadvantaged backgrounds (relative to White teachers who are high scoring), this might bias downward our estimate of the interaction between teacher race and MTEL performance. In other words, such sorting would tend to bias our estimates in favor of falsely detecting bias in the predictive validity of the MTEL. Similarly, if the direction of sorting were reversed, we may not detect bias that actually exists.

We test for nonrandom student sorting in **Table 6** by restricting the sample to teachers in Grade 5 or higher and including twice-lagged student achievement in our value-added estimates.[15] Rothstein (2009) finds that the inclusion of twice-lagged scores reduces bias due to nonrandom teaching assignments within schools. A significant movement in the coefficient on the interaction between MTEL and teachers of color resulting from the inclusion of twice-lagged scores would therefore indicate bias due to nonrandom sorting.

We find little evidence of bias from nonrandom student sorting. The coefficients on MTEL scores for the full sample are nearly identical to the coefficients in Table 5. The interaction effects are somewhat more positive for performance ratings and somewhat more negative for

[15] For the models using teacher performance ratings, we additionally include lagged test scores, which are omitted from the main analyses that use data on students in early elementary grades as well.

ELA value added. But the interactions remain statistically insignificant in math and ELA value added and statistically significant for teacher evaluations.

5.3 Sample Selection Bias

Ideally, we would estimate the relationship between MTEL scores and performance measures for all teacher candidates who take the MTEL. However, this is clearly infeasible, given that only a fraction of MTEL test takers are observed in Massachusetts public schools between 2014 and 2019 (performance ratings) or in tested grades and subjects between 2012 and 2019 (value added). In particular, as we discuss below, we are relatively unlikely to observe candidates who fail the MTEL on their first attempt. This kind of sample selection may bias our estimates of the relationships between MTEL scores and teacher effectiveness and our comparisons of predictive validity for White candidates and candidates of color.

In our estimates of the overall relationship between MTEL and teacher performance, the primary concern is that teachers who initially fail the test but still end up in the teaching profession differ systematically from those candidates who fail and never enter the workforce. For instance, more motivated candidates may be more likely to retake the MTEL and also may ultimately be more effective teachers. This trend would tend to bias our estimates of the relationship between MTEL and teacher performance downward. But other factors (e.g., financial resources) could also influence candidates' abilities to retake the MTEL, so it is unclear which direction this bias would operate.

In the standard setup, two general factors govern the sample selection bias (Heckman, 1979). The first factor is the strength of the relationship between MTEL scores and entry into the teaching profession. In our context, if the relationship between MTEL scores and employment is strong, then our estimates of the relationship between MTEL scores and teacher effectiveness measures are likely attenuated toward zero. Intuitively, the more MTEL predicts the likelihood of employment, the less representative the final sample is of all test takers. The second factor is the strength of the relationship between (unobserved) teacher effectiveness and the likelihood of teacher employment. The intuition in this case is similar: The more *potential* teacher effectiveness (i.e., any factors that influence teacher effectiveness that are not observed in our data) predicts employment, the less variation in effectiveness exists among hired teachers, and the weaker the relationship between preservice measures and observed teacher effectiveness.[16] When making comparisons in the predictive validity of MTEL by candidate race, we are concerned about differences in the strength of these relationships between White candidates and candidates of color.

In **Appendix Figure A.1** and **Table 7**, we demonstrate some evidence of potential differential selection by teacher race/ethnicity that motivate our approach for assessing the potential for

[16] Although this second factor may theoretically attenuate our estimates, existing evidence suggests that employers have little information about teachers' potential effectiveness before they are hired and that even potential signals (e.g., interviews, sample lessons, recommendation letters) are not strongly correlated with employment offers (Jacob et al., 2018). We therefore view this source of bias as less important in practice.

selection bias. First, in Appendix Figure A.1, we plot the likelihood that a teacher is observed in the value added and performance rating samples against average first-time MTEL scores. We plot results separately for teachers who pass both required CLSTs on the first attempt (right panel) and for teachers who initially fail one or both of the tests (left panel). In all cases, the likelihood of entering the analytical sample increases sharply with average scores among teachers initially failing one of the tests. But the relationship between sample inclusion and test scores is much weaker among candidates who pass both of the required tests. This finding is consistent with the findings of Boyd et al. (2013), who demonstrate that schools primarily value whether a teacher has obtained their teaching qualifications and do not highly value higher levels of licensure test performance. The relatively flat relationship between MTEL scores and sample selection above the passing threshold suggests that any sample selection bias among teachers who pass all of their tests on the first attempt should be relatively small because the relationship between MTEL and sample inclusion is weak (the first factor discussed above). In a standard selection model (Heckman, 1979), the bias is the product of these two factors. Thus, if the relationship between MTEL scores and employment is weak, we would not expect a significant selection bias.

An additional concern in our context is that the relationship between MTEL and employment for candidates below the passing threshold appears to differ for teacher candidates of color. We explore this pattern in more detail in Table 7. Using data on all first-time test takers in Massachusetts, we estimate linear probability models where the dependent variable is an indicator that a teacher retakes (or passes) the MTEL within 53 weeks, and we include the same set of interactions as in Equation (4).[17]

$$R_{jt} = C_j\alpha + MTEL_j\beta + MTEL_j * C_j\gamma + \lambda_t + \eta_{jt} \qquad (5)$$

We additionally include year fixed effects λ_t. We estimate Equation (5) for each of the required Communication and Literacy Skills tests and for all subject tests.

In Table 7, we show that, conditional on initial MTEL score, teachers of color are about 7 to 13 percentage points less likely to retake the test within one year (column 1), a finding which is consistent with Rucinski and Goodman (2019). The deterrent effect is larger for the CLSTs, especially the reading test, than the subject tests. In addition, teachers of color are more sensitive to the initial test score. That is, the discrepancy in retest rates increases as candidates get further from the passing score. In columns 3 and 4, we replace the retake outcome with an indicator for passing the MTEL. Candidates of color are subsequently less likely to pass the MTEL (column 3). In column 4, we include an indicator for any retake in the regression. Comparing the coefficient on candidate of color in columns 3 and 4, we see that differences in retake rates (conditional on initial MTEL score) explain about half of the difference in eventual pass rates. These findings are at least suggestive that the MTEL presents an additional barrier to candidates

[17] We observe a significant mass point of teachers retaking the MTEL at 371 days (53 weeks), so we include these teachers retaking the MTEL in our analysis.

of color. Even adjusting for initial test scores, candidates of color are less likely to attempt and less likely to pass the licensure testing requirement.

Given the evidence of discrepancies in retesting rates, we focus, in **Table 8,** on the sample of teacher candidates who pass both required CLSTs on the first attempt. First-time passers comprise about 75% of all teachers in our sample of test takers and about 60% of all teachers of color. In general, the results are similar to the baseline results in Table 5, and there is no clear pattern across performance measures. Among all teachers, the relationships between MTEL scores and math value added are nearly identical to the relationships estimated in Table 5. The relationships between MTEL scores and ELA performance are somewhat weaker and not statistically significant in some specifications, but they are similar or stronger for the teacher performance ratings.

We generally find stronger relationships between MTEL scores and teacher performance measures for candidates of color in columns 4–6. In both math and ELA, we estimate interactions between teachers of color and MTEL scores that are mostly positive; a caution, however, is that these interactions are estimated imprecisely and neither the results in math or ELA are statistically significant. For teacher performance ratings, the interactions are more positive than in Table 5 and are nearly the same magnitude as the overall relationships between MTEL scores and performance ratings. Although the estimates are imprecise, so we cannot be very definitive, these findings generally support the conclusion that sample selection associated with teacher entry into the workforce is not a significant concern.

One additional concern is that teachers may leave the workforce at different rates. This differential attrition, if it exists, should not substantially affect our estimates given that we control for teacher experience and implicitly use averages of the outcomes at the teacher level as the dependent variable. However, we repeat the main analysis among novice teachers in **Appendix Table A.4**. Although the estimates are less precise, we find the same patterns as in Table 5. In particular, we find little evidence of differential predictive validity in math or ELA value added by teacher race/ethnicity.[18]

6. Conclusion

In this study, we assess the relationship between licensure tests of teacher basic skills and content knowledge and their effectiveness as classroom teachers. The overall relationships between MTEL scores and value added are comparable to what has been found in other states. For example, the estimated relationships between MTEL scores and math value added in Table 5 all fall within the range of estimates for similar models estimated for WEST-B and WEST-E licensure tests in the state of Washington (Goldhaber et al., 2017). We also demonstrate significant, positive relationships between licensure test scores and in-service performance

[18] Results are also robust to considering whether candidates pass the test on the first attempt (see Appendix Table A.5). And, as shown in Appendix Table A.6, these results are also robust to using alternative measures of teacher race/ethnicity measured from the EPIMS data discussed in Section 3.2

ratings.

We do not find that licensure tests are less predictive of teacher performance for teachers of color in Massachusetts. Instead, we find that correlations between test scores and teacher value-added are similar across racial groups, and that tests are somewhat *more* strongly correlated with summative performance ratings among teachers of color. Nonetheless, our findings are not dispositive about whether testing requirements disproportionately exclude potentially effective candidates of color. First, we find that candidates of color are less likely to retake the MTEL after they fail on the first attempt than white teachers with similar scores. Furthermore, to the extent that licensure tests are noisy measures of potential performance, it is unlikely that the differences in teacher test performance are fully explained by differences in true teaching potential (Jencks, 1998).

Our findings on the (lack of) differential predictive validity of MTEL scores for candidates of color relative to White candidates differ from the only prior study that considers this specific issue, Goldhaber and Hansen (2010). This prior study found evidence that scores on some Praxis tests in North Carolina are better predictors of value added for White than Black teachers. There are a number of potential explanations for these differences in findings. The different findings could, for instance, be related to the licensure tests themselves (i.e., between the Praxis and MTEL); for instance, Massachusetts engages a bias review panel as part of the test development process. There are also a number of factors that could influence the populations of tested teacher candidates and teachers in the workforce in each state, such as differences in the populations pursuing teaching; differences in the thresholds established for passing the tests; or the patterns of test retakes and workforce entry between states.

That said, it is clear from this analysis that licensure tests play a critical role in determining employment eligibility and have disparate effects on eligibility by teacher candidate race/ethnicity. Given the policy context in which courts have come down on different sides of lawsuits challenging testing requirements (e.g., Gulino v. Board of Education, 2002; Mitchell et al., 2001), we believe it is important to replicate the analyses we have conducted in Massachusetts in other states and with other types of licensure tests. If the difference in findings between Massachusetts and North Carolina reflects the particular bias mitigation strategies used in Massachusetts, then these policies may be important to ensure that licensure tests are equally predictive of performance for all teachers. On the other hand, if licensure tests do consistently measure a latent teaching skill, then policymakers concerned about the diversity of the teaching profession may wish to broaden the range of teaching skills assessed by licensure systems. Some scholars have pointed to the potential of newer assessments that rely on more direct evidence of effective classroom teaching, such as the edTPA, as potentially having less impact on workforce diversity (e.g., Gershenson et al., 2021). While the edTPA has also been shown to have significant disparities by teacher race/ethnicity (edTPA, 2018; Goldhaber et al., 2017), the gaps

in performance are generally smaller in magnitude than what we document for MTEL.[19] In addition, there is currently limited direct evidence on the effects of licensure policies on the effectiveness or diversity of the teaching profession. Additional research would help policymakers better understand how policies that seek to enforce professional standards might impact the diversity of the teacher workforce.

[19] Goldhaber et al. (2017) find that Hispanic candidates performed about 0.5 standard deviations lower on the edTPA in Washington State than White candidates, while edTPA (2018) reports from national data that Black candidates score on average about 0.4 standard deviations lower on the edTPA than White candidates.

References

Angrist, J. D., & Guryan, J. (2004). Teacher testing, teacher education, and teacher characteristics. *American Economic Review: Papers and Proceedings, 94*(2), 241–246.

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

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*. https://doi.org/10.1016/j.econedurev.2019.101919

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review, 68,* 89–103.

Backes, B., & Hansen, M. (2018). The impact of Teach For America on non-test academic outcomes. *Education Finance and Policy, 13*(2), 168–193.

Baker, S. (1995). Testing equality: The National Teacher Examination and the NAACP's legal campaign to equalize teachers' salaries in the south, 1936-63. *History of Education Quarterly, 35*(1), 49–64.

Ballou, D., & Podgursky, M. J. (1998). Teacher recruitment and retention in public and private schools. *Journal of Policy Analysis and Management, 17*(3), 393–417.

Bonhomme, S., Lamadon, T., & Manresa, E. (2019). *Discretizing unobserved heterogeneity*. Unpublished manuscript.

Bonhomme, S., & Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica, 83*(3), 1147–1184.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. *The Future of Children*, *17*(1), 45–68. https://doi.org/10.1353/foc.2007.0000

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Analyzing the determinants of the matching of public school teachers to jobs: Disentangling the preferences of teachers and employers. *Journal of Labor Economics, 31*(1), 83–117.

Center for Constitutional Rights. (2012). *Federal court rules against NYC board of ed: Teacher exam discriminated*. https://ccrjustice.org/node/1045

Chen, B., Cowan, J., Goldhaber, D., & Theobald, R. (2019). From the clinical experience to the classroom: Assessing the predictive validity of the Massachusetts Candidate Assessment of Performance. CALDER Working Paper 221-0819.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

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.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources, 45*(3), 655–681. https://doi.org/10.3368/jhr.45.3.655

Compton, E., Bentley, M., Ennis, S., & Rastogi, S. (2013). *2010 census race and Hispanic origin alternative questionnaire experiment*. United States Census Bureau. https://www.census.gov/2010census/pdf/2010_Census_Race_HO_AQE.pdf

Cowan, J., Goldhaber, D., Hayes, K., & Theobald, R. (2017). Missing elements in the discussion of teacher shortages. *Educational Researcher, 45*(8), 460–462.

Cowan, J., Goldhaber, D., & Theobald, R. (2018). *Massachusetts teacher preparation and licensure: Performance Review Program for Initial Licensure study*. American Institutes for Research.

Cowan, J., Goldhaber, D., & Theobald, R. (2020). *Performance evaluations as a measure of teacher effectiveness when standards differ: Accounting for variation across classrooms, schools, and districts*. National Center for the Analysis of Longitudinal Data in Education Research.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics, 86*(1), 195–210.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review,* 158–165.

Dee, T. S., & Domingue, B. W. (2019). *Did a "traumatic" test question create racial bias?* Stanford Center for Education Policy Analysis.

Drake, S., Auletto, A., & Cowen, J. M. (2019). Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal*, 56(5), 1800-1833.

edTPA (2018). Educative assessment and meaningful support: 2017 edTPA administrative report. Stanford University.

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review, 45,* 44–52.

Gershenson, S., Hansen, M., & Lindsay, C. (2021). *Teacher diversity and student success: Why racial representation matters in the classroom.*

Gershenson, S., Hart, C. M. D., Hyman, J., Lindsay, C., & Papageorge, N. W. (2018). *The long-run impacts of same-race teachers* (Working Paper No. 25254). National Bureau of Economic

Research. https://doi.org/10.3386/w25254

Goldhaber, D. (2004). Why do we license teachers? In F. M. Hess, A. J. Rotherham, & K. Walsh (Eds.), *A qualified teacher in every classroom?* Cambridge: Harvard Education Press.

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

Goldhaber, D. (2011). Licensure: Exploring the value of this gateway to the teacher workforce. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 315–339). Elsevier B.V.

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

Goldhaber, D., Gratz, T., & Theobald, R. (2017). What's in a teacher test? Assessing the relationship between teacher licensure test scores and student STEM achievement and course-taking. *Economics of Education Review*, *61*, 112–129. https://doi.org/10.1016/j.econedurev.2017.09.002

Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal, 47*(1), 218–251. https://doi.org/10.3102/0002831209348970

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.

Gulino v. Board of Educ., City of New York, 236 F. Supp. 2d 314 (S.D.N.Y. 2002)

Harris, E. A. (2015, June 17). Tough tests for teachers, with question of bias. *The New York Times*. https://www.nytimes.com/2015/06/18/nyregion/with-tougher-teacher-licensing-exams-a-question-of-racial-discrimination.html

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.

Hendricks, M. (2015). Public Schools Are Hemorrhaging Talented Teachers: Can Higher Salaries Function as a Tourniquet?. Available at SSRN 2564703.

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

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *The Journal of Economic Perspectives, 30*(3), 85–108.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics, 166*, 81–97.

Jencks, C. (1998). Racial bias in testing. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 55–85). Brookings Institution Press.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers?* Bill and Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources, 54*(1), 1–36.

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.

Lambert, D. (2020, January 23). *California moves closer to eliminating, replacing reading instruction test that has blocked thousands from teaching credential*. EdSource. https://edsource.org/2020/california-moves-closer-to-eliminating-replacing-reading-instruction-test-that-has-blocked-thousands-from-teaching-credential/622830

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (Eds.). (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. National Academy Press.

Nettles, M. T., Scatton, L. H., Steinberg, J. H., & Tyler, L. L. (2011). *Performance and passing rate differences of African American and White prospective teachers on Praxis™ examinations* (ETS RR-11-08). NEA and ETS. https://www.ets.org/Media/Research/pdf/RR-11-08.pdf

Office of Management and Budget. (1997). Revisions to the standards for the classification of federal data on race and ethnicity. Federal Register, 62(210), 58782-58790.

Partelow, L., Spong, A., Brown, C., & Johnson, S. (2017, September 14). *America needs more teachers of color and a more selective teaching profession.* Center for American Progress. https://www.americanprogress.org/issues/education-k-12/reports/2017/09/14/437667/america-needs-teachers-color-selective-teaching-profession/

Petchauer, E. (2019, May 7). We need more teachers of color. Let's scrap exams that keep them out of the classroom. *Education Week.* https://www.edweek.org/tm/articles/2019/05/07/we-need-more-teachers-of-color-lets.html

Ravitch, D. (2003, August 23). *Diane Ravitch, Ph.D.—A brief history of teacher professionalism—White house conference on preparing tomorrow's teachers* [Speeches and Testimony; Conference Papers/Proceedings]. U.S. Department of Education. https://www2.ed.gov/admins/tchrqual/learn/preparingteachersconference/ravitch.html

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

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537–571.

Rucinski, M., & Goodman, J. (2019) *Racial diversity in the teacher pipeline: Evidence from Massachusetts.* Harvard Kennedy School Policy Brief.

Sass, T. R. (2015). Licensure and worker quality: A comparison of alternative routes to teaching. *The Journal of Law and Economics*, 58(1), 1-35.

Skinner, K., Betz, K. D., & Wright, A. (2020, April 23). *Suspension of teacher license test amid COVID-19 crisis likely to 'open up some doors' for potential educators*. Mississippi Today. https://mississippitoday.org/2020/04/23/suspension-of-teacher-license-test-amid-covid-19-crisis-likely-to-open-up-some-doors-for-potential-educators/

Sleeter, C. E. (2017). Critical race theory and the whiteness of teacher education. *Urban Education, 52*(2), 155–169.

Smith, G. P. (1988). Tomorrow's White teachers: A response to the Holmes Group. *The Journal of Negro Education, 57*(2), 178–194.

Tillman, L. C. (2004). (Un)intended consequences? The impact of the *Brown v. Board of Education* decision on the employment status of black educators. *Education and Urban Society, 36*(3): 280-303.

Tyler, L., Whiting, B., Ferguson, S., Eubanks, S., Steinberg, J., Scatton , L., & Bassett, K. (2011). *Toward increasing teacher diversity: Targeting support and intervention for teacher licensure candidates*. NEA and ETS. https://www.ets.org/Media/Research/pdf/ETS-NEA-2011-01.pdf

U.S. Department of Education (2017), National Center for Education Statistics, Schools and Staffing Survey (SASS), National Teacher and Principal Survey (NTPS), "Public School Teacher Data File," 2015-16.

Vars, F. E., & Bowen, W. G. (1998). *Scholastic Aptitude Test scores, race, and academic performance in selective colleges and universities.* Brookings Institute.

Weir, K. (2016). Inequality at school. *Monitor on Psychology*, *47*(10). http://www.apa.org/monitor/2016/11/cover-inequality-school

Tables and Figures

Figure 1. MTEL Scores and Math Value Added

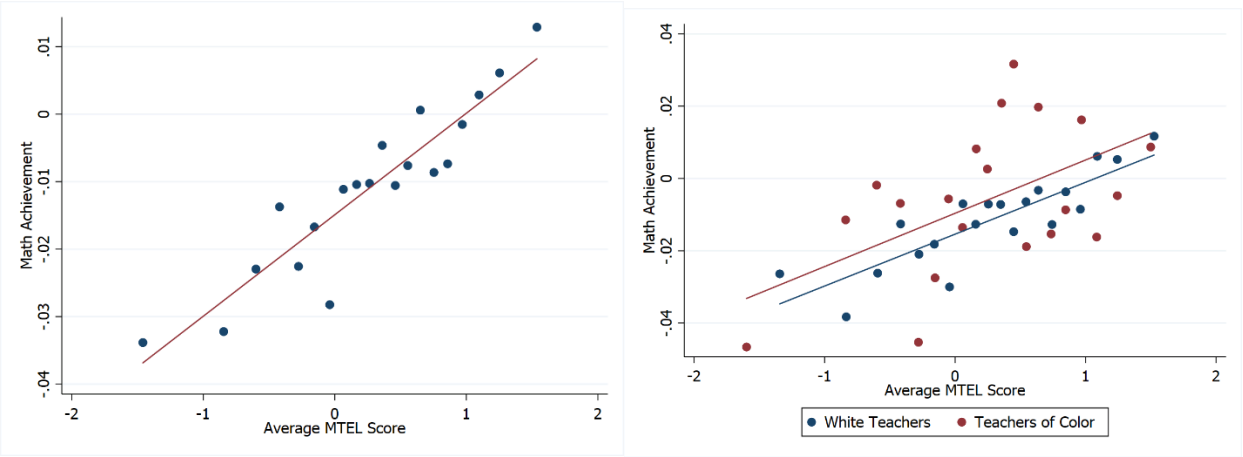


Figure 2. MTEL Scores and ELA Value Added

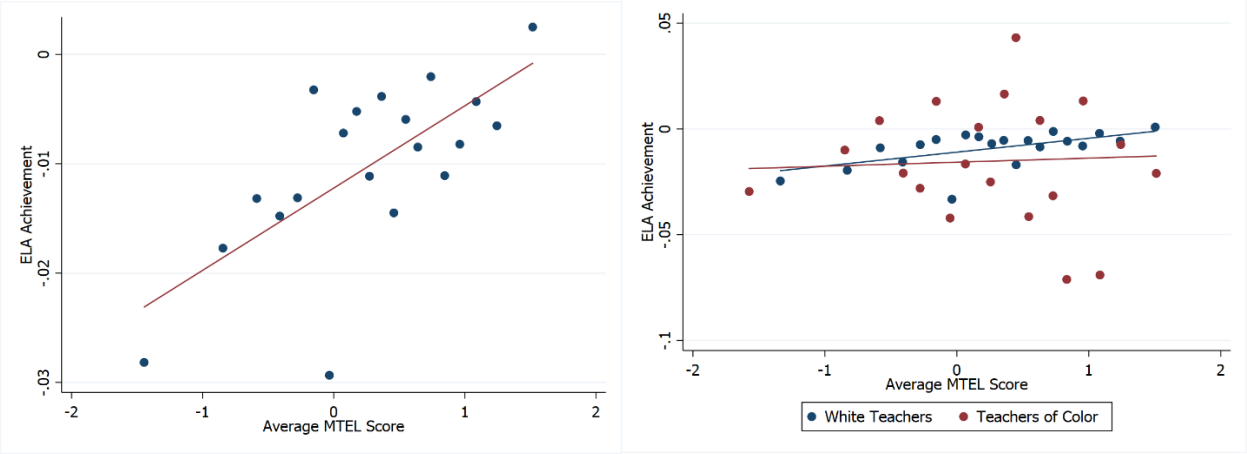
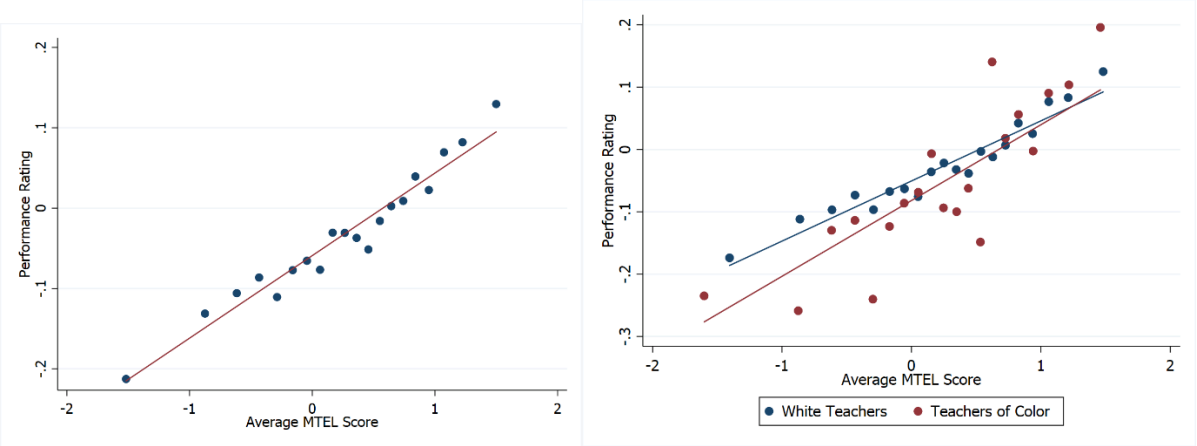


Figure 3. MTEL Scores and Teacher Performance Ratings



Notes: Plots of teacher effectiveness measures and average first-time MTEL scores. Value-added models calculated using grouped fixed effects methods described in the text. Bins represent 20 equally sized groupings by MTEL scores.

Table 1: First MTEL Scores and Passing Status by Candidate Race/Ethnicity

	(1) American Indian	(2) Asian	(3) Black	(4) Hispanic	(5) Other	(6) White
Panel A. CLST Results						
Pass	68.8	70.8	49.5	54.4	67.6	76.6
Standardized Score	0.04	0.05	-0.46	-0.41	0.01	0.22
N	292	3,926	5,304	5,808	3,805	139,051
Panel B. Subject Test Results						
Pass	61.4	74.3	39.9	60.5	63.4	69.3
Standardized Score	-0.06	0.28	-0.64	-0.07	0.04	0.22
N	233	3,429	4,015	4,853	3,237	114,909

Notes: MTEL scores by teacher race/ethnicity in licensure testing. Observations are at the candidate level. We use the mode of a teacher's reported race/ethnicity codes as the assigned identity. Excludes teachers without reported race/ethnicity in the testing data.

Table 2: Summary Statistics (Math Value-Added Sample)

	All Teachers	White Teachers	Teachers of Color	Missing Race/Ethnicity
Math score	-0.006	0.006	-0.198	0.065
	(0.922)	(0.915)	(0.968)	(0.935)
Lagged math score	-0.023	-0.010	-0.225	0.042
	(0.920)	(0.913)	(0.962)	(0.928)
Lagged ELA score	-0.033	-0.017	-0.261	0.021
	(0.927)	(0.921)	(0.958)	(0.931)
English learner	0.054	0.048	0.122	0.055
	(0.227)	(0.214)	(0.327)	(0.227)
Male student	0.504	0.504	0.501	0.503
	(0.500)	(0.500)	(0.500)	(0.500)
Student eligible for FRL	0.365	0.348	0.570	0.344
	(0.481)	(0.476)	(0.495)	(0.475)
Full-inclusion SWD	0.125	0.126	0.121	0.124
	(0.331)	(0.332)	(0.326)	(0.329)
Partial-inclusion SWD	0.022	0.022	0.025	0.022
	(0.148)	(0.147)	(0.157)	(0.145)
Asian students	0.063	0.062	0.069	0.073
	(0.243)	(0.241)	(0.253)	(0.260)
Black students	0.083	0.072	0.198	0.086
	(0.276)	(0.259)	(0.398)	(0.280)
Hispanic students	0.177	0.164	0.332	0.169
	(0.382)	(0.370)	(0.471)	(0.375)
Experience	8.250	8.145	7.005	10.949
	(5.859)	(5.638)	(5.517)	(7.824)
Teacher of color	0.082			
	(0.275)			
Average first-time MTEL scores	0.273	0.294	-0.059	0.381
	(0.725)	(0.698)	(0.812)	(0.840)
Passed all tests on first attempt	0.492	0.491	0.349	0.665
	(0.500)	(0.500)	(0.477)	(0.472)
Observations	2,002,904	1,712,646	153,235	137,023
Unique Teachers	13,957	11,671	1,264	1,022

Table 3. Summary Statistics (ELA Value-Added Sample)

	All Teachers	White Teachers	Teachers of Color	Missing Race/Ethnicity
ELA score	0.002	0.014	-0.180	0.018
	(0.914)	(0.911)	(0.942)	(0.914)
Lagged math score	-0.007	0.002	-0.181	0.025
	(0.921)	(0.917)	(0.958)	(0.925)
Lagged ELA score	-0.012	-0.001	-0.201	0.012
	(0.921)	(0.917)	(0.953)	(0.924)
English learner	0.050	0.045	0.109	0.054
	(0.218)	(0.208)	(0.312)	(0.225)
Male student	0.503	0.503	0.497	0.502
	(0.500)	(0.500)	(0.500)	(0.500)
Student eligible for FRPL	0.358	0.346	0.555	0.345
	(0.480)	(0.476)	(0.497)	(0.475)
Full-inclusion SWD	0.124	0.124	0.115	0.125
	(0.329)	(0.330)	(0.319)	(0.331)
Partial-inclusion SWD	0.021	0.021	0.024	0.021
	(0.143)	(0.142)	(0.154)	(0.142)
Asian students	0.064	0.063	0.072	0.073
	(0.245)	(0.243)	(0.258)	(0.260)
Black students	0.079	0.071	0.201	0.076
	(0.270)	(0.257)	(0.401)	(0.265)
Hispanic students	0.174	0.165	0.308	0.166
	(0.379)	(0.371)	(0.462)	(0.372)
Experience	8.059	7.861	7.083	10.858
	(5.869)	(5.637)	(5.792)	(7.351)
Teacher of color	0.066			
	(0.249)			
Average first-time MTEL scores	0.282	0.290	0.030	0.386
	(0.702)	(0.684)	(0.785)	(0.775)
Passed all tests on first attempt	0.551	0.550	0.379	0.695
	(0.497)	(0.497)	(0.485)	(0.461)
Observations	1,880,312	1,611,439	114,667	154,206
Unique teachers	13,978	11,817	1,091	1,070

Table 4. Summary Statistics (Performance Ratings Sample)

	All Teachers	White Teachers	Teachers of Color	Missing Race/Ethnicity
Class average lagged ELA score	-0.117	-0.097	-0.405	-0.053
	(0.646)	(0.629)	(0.723)	(0.662)
Class average lagged math score	-0.103	-0.086	-0.363	-0.034
	(0.661)	(0.647)	(0.725)	(0.674)
Limited English proficient students	0.101	0.088	0.220	0.111
	(0.207)	(0.186)	(0.310)	(0.231)
Male students	0.509	0.509	0.509	0.508
	(0.125)	(0.124)	(0.128)	(0.128)
FRPL students	0.388	0.370	0.585	0.369
	(0.326)	(0.319)	(0.330)	(0.324)
Full-inclusion SWD	0.116	0.117	0.110	0.108
	(0.133)	(0.134)	(0.130)	(0.129)
Partial-inclusion SWD	0.025	0.025	0.025	0.024
	(0.072)	(0.073)	(0.071)	(0.067)
Asian students	0.066	0.064	0.073	0.073
	(0.112)	(0.107)	(0.139)	(0.126)
Black students	0.095	0.084	0.211	0.091
	(0.159)	(0.144)	(0.239)	(0.157)
Hispanic students	0.202	0.188	0.357	0.192
	(0.252)	(0.241)	(0.302)	(0.251)
Experience	8.619	8.262	7.790	12.174
	(6.058)	(5.612)	(6.275)	(7.890)
Teacher of color	0.092			
	(0.290)			
Average first-time MTEL scores	0.266	0.291	-0.059	0.331
	(0.731)	(0.692)	(0.845)	(0.859)
Passed all tests on first attempt	0.523	0.522	0.350	0.665
	(0.499)	(0.499)	(0.477)	(0.472)
Observations	852,087	694,949	70,830	86,308
Unique teachers	53,613	43,345	5,337	4,931

Table 5. Licensure Tests and Teacher Performance

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Value Added						
MTEL	0.024***	0.015***	0.012***	0.024***	0.014***	0.011***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
MTEL x TOC				0.004	0.000	0.002
				(0.007)	(0.006)	(0.006)
N	2,002,904	2,002,904	2,002,904	1,865,881	1,865,881	1,865,881
Panel B. ELA Value Added						
MTEL	0.014***	0.008***	0.005***	0.014***	0.007***	0.003*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
MTEL x TOC				-0.005	-0.005	-0.003
				(0.007)	(0.006)	(0.006)
N	1,880,311	1,880,311	1,880,311	1,726,105	1,726,105	1,726,105
Panel C. Teacher Performance Evaluations						
MTEL	0.104***	0.103***	0.081***	0.097***	0.096***	0.072***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)
MTEL x TOC				0.029*	0.025*	0.026*
				(0.015)	(0.015)	(0.013)
N	776,806	776,806	776,169	698,422	698,422	697,843
School-Grade Group FE		Y			Y	
School-Grade FE			Y			Y

Notes: Regressions of teacher value added on average first-time MTEL scores. Covariates in value-added regressions are described in the text. Regressions are weighted by the inverse of the number of observations for each teacher. Standard errors clustered by teacher in parentheses. * $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$.

Table 6. Licensure Tests and Teacher Performance With Twice-Lagged Achievement

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Value Added						
MTEL	0.022***	0.014***	0.011***	0.015***	0.005**	0.003
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
MTEL x TOC				0.007	0.001	0.001
				(0.008)	(0.007)	(0.007)
N	1,561,981	1,561,981	1,561,981	1,457,967	1,457,967	1,457,967
Panel B. ELA Value Added						
MTEL	0.015***	0.009***	0.005**	0.008***	0.000	-0.004*
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
MTEL x TOC				-0.007	-0.008	-0.007
				(0.008)	(0.007)	(0.007)
N	1,467,486	1,467,486	1,467,486	1,343,900	1,343,900	1,343,900
Panel C. Teacher Performance Ratings						
MTEL	0.113***	0.111***	0.084***	0.100***	0.099***	0.069***
	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.005)
MTEL x TOC				0.049***	0.045**	0.044***
				(0.019)	(0.019)	(0.016)
N	543,824	543,824	543,482	485,890	485,890	485,577
School-Grade Group FE		Y			Y	
School-Grade FE			Y			Y

Notes: Regressions of teacher value added on average first-time MTEL scores (Grade 5 and higher). Covariates in value-added regressions are described in the text. Regressions are weighted by the inverse of the number of observations for each teacher. Standard errors clustered by teacher in parentheses. * $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$.

Table 7. Probability of Retaking or Passing MTEL, by Race and Initial Score

	(1)	(2)	(3)	(4)
Outcome:	Retake	Retake	Pass	Pass
Panel A. CLST Reading				
MTEL	0.051***	0.038***	0.130***	0.091***
	(0.004)	(0.004)	(0.004)	(0.004)
Teacher of Color	-0.128***	-0.083***	-0.177***	-0.081***
	(0.008)	(0.012)	(0.008)	(0.007)
MTEL x TOC		0.043***		
		(0.009)		
Retake				0.756***
				(0.004)
N	22,168	22,168	22,168	22,168
Panel B. CLST Writing				
MTEL	0.075***	0.063***	0.190***	0.138***
	(0.004)	(0.005)	(0.004)	(0.003)
Teacher of Color	-0.087***	-0.056***	-0.120***	-0.060***
	(0.007)	(0.010)	(0.007)	(0.006)
MTEL x TOC		0.032***		
		(0.008)		
Retake				0.694***
				(0.004)
N	27,636	27,636	27,636	27,636
Panel C. MTEL Subject Tests				
MTEL	0.099***	0.094***	0.243***	0.182***
	(0.003)	(0.004)	(0.003)	(0.003)
Teacher of Color	-0.073***	-0.052***	-0.103***	-0.058***
	(0.006)	(0.010)	(0.007)	(0.006)
MTEL x TOC		0.020**		
		(0.008)		
Retake				0.625***
				(0.003)
N	36,772	36,772	36,772	36,772

Notes: Estimates from linear probability models of retaking or passing the given MTEL test within 53 weeks of initial test. Retake indicates that the candidate attempted the same test type at least once within 53 weeks of initial test date. Pass indicates that the candidate passed the same test type within 53 weeks of initial test date. Sample includes all candidates who initially failed the given test. All models control for year fixed effects. Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

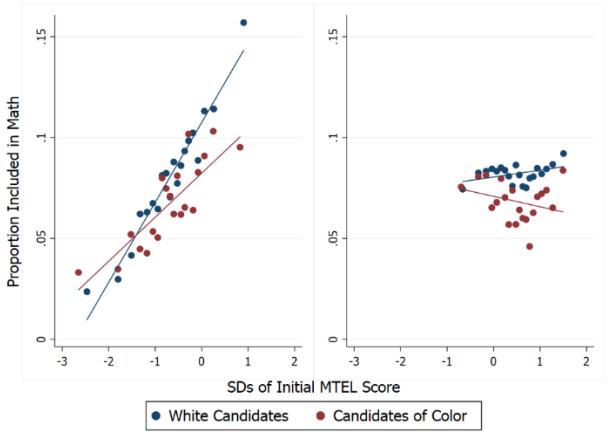
Table 8. Licensure Tests and Teacher Performance (First-Time Passing Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Value-Added						
MTEL	0.028***	0.018***	0.015***	0.027***	0.016***	0.013***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
MTEL x TOC				0.015	0.009	0.011
				(0.013)	(0.011)	(0.010)
N	1,398,383	13,98,383	1,398,383	1,328,428	1,328,428	1,328,428
Panel B. ELA Value-Added						
MTEL	0.012***	0.005*	0.001	0.010***	0.004	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
MTEL x TOC				0.006	-0.002	0.004
				(0.012)	(0.011)	(0.010)
N	1,405,708	1,405,708	1,405,708	1,317,958	1,317,958	1,317,958
Panel C. Performance Ratings						
MTEL	0.121***	0.099***	0.094***	0.115***	0.092***	0.088***
	(0.007)	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)
MTEL x TOC				0.074***	0.074***	0.062***
				(0.027)	(0.024)	(0.023)
N	562,506	562,506	562,052	522,134	522,134	521,716
School-grade group FE		X			X	
School-grade FE			X			X

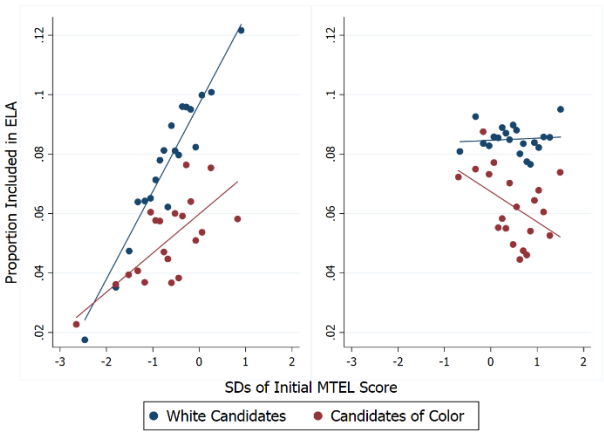
Notes: Regressions of teacher value added on average first-time MTEL scores (first-time passers only). Covariates in value-added regressions are described in the text. Regressions are weighted by the inverse of the number of observations for each teacher. Standard errors clustered by teacher in parentheses. * $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$.

Appendix A. Additional Results

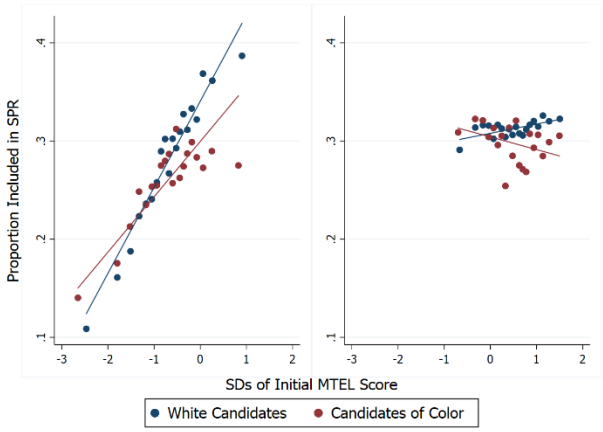
Appendix Figure A.1. Average MTEL Scores and Sample Inclusion



(a) Math



(b) ELA



(c) Performance Ratings

Appendix Table A.1. Teacher Race/Ethnicity Reported in MTEL and EPIMS Data

EPIMS / MTEL	White	Hispanic	Black	Asian	Am. Indian	Multiple Races	Pacific Islander
White	464,088	1,221	353	120	158	708	52
Hispanic	2,995	12,255	121	15	11	26	0
Black	515	136	12,647	15	24	379	4
Asian/Pacific Islander	1,009	79	33	7,650	61	430	187
American Indian	475	27	24	0	220	108	9
Other	6,819	1,158	1,724	538	36	901	12

Notes: Teacher race/ethnicity in licensure testing and administrative data sets. Sample consists of all test takers linked to administrative teacher records. Observations are at the teacher-school-year level. In both data sets, we use the mode of a teacher's reported race/ethnicity codes as the assigned identity. Excludes teachers without reported race/ethnicity in the testing data.

Appendix Table A.2. Licensure Tests and Teacher Performance (Highest Score)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Value Added Sample						
MTEL	0.032***	0.020***	0.017***	0.031***	0.019***	0.015***
	(0.003)	(0.003)	(0.002)	(0.004)	(0.003)	(0.003)
MTEL x TOC				0.014	0.008	0.009
				(0.010)	(0.008)	(0.009)
N	2,002,904	2,002,904	2,002,904	1,865,881	1,865,881	1,865,881
Panel B. ELA Value Added Sample						
MTEL	0.019***	0.011***	0.007***	0.018***	0.009***	0.004
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
MTEL x TOC				-0.001	-0.001	0.001
				(0.009)	(0.008)	(0.008)
N	1,880,311	1,880,311	1,880,311	1,726,105	1,726,105	1,726,105
Panel C. Teacher Performance Ratings						
MTEL	0.147***	0.145***	0.114***	0.140***	0.139***	0.103***
	(0.006)	(0.006)	(0.005)	(0.007)	(0.007)	(0.006)
MTEL x TOC				0.061***	0.056**	0.049***
				(0.022)	(0.022)	(0.019)
N	776,806	776,806	776,169	698,422	698,422	697,843
School-Grade Group FE		Y			Y	
School-Grade FE			Y			Y

Appendix Table A.3. MTEL and Teacher Performance by Test Type

	Math		ELA		SPR	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. CLST Reading						
MTEL	0.012***	0.007***	0.008***	0.003*	0.043***	0.031***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)
MTEL x TOC	-0.003	-0.004	-0.009	-0.004	0.018	0.018
	(0.006)	(0.005)	(0.006)	(0.005)	(0.013)	(0.011)
N	1,747,751	1,747,751	1,650,063	1,650,063	680,637	680,637
Panel B. CLST Writing						
MTEL	0.015***	0.008***	0.009***	0.004**	0.072***	0.059***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)
MTEL x TOC	0.001	-0.004	-0.007	-0.008	0.004	0.006
	(0.006)	(0.005)	(0.006)	(0.006)	(0.012)	(0.011)
N	1,747,005	1,747,005	1,648,417	1,648,417	680,231	680,231
Panel C. Subject Tests						
MTEL	0.016***	0.011***	0.007***	0.004**	0.055***	0.043***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.004)	(0.003)
MTEL x TOC	0.002	-0.000	0.002	-0.000	0.021*	0.022**
	(0.005)	(0.005)	(0.006)	(0.005)	(0.011)	(0.010)
N	1,864,727	1,864,727	1,723,844	1,723,844	709,374	709,374
School-grade group FE		X		X		

Appendix Table A.4. Licensure Tests and Teacher Performance (Novices)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Value Added						
MTEL	0.033***	0.022***	0.018***	0.031***	0.019***	0.015***
	(0.006)	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)
MTEL x TOC				0.010	0.010	0.009
				(0.018)	(0.015)	(0.015)
N	116,490	116,490	116,490	109,935	109,935	109,935
Panel B. ELA Value Added						
MTEL	0.016***	0.005	0.003	0.016***	0.004	0.002
	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)
MTEL x TOC				-0.014	-0.016	-0.013
				(0.017)	(0.015)	(0.014)
N	118,208	118,208	118,208	111,552	111,552	111,552
Panel C. Teacher Performance Ratings						
MTEL	0.098***	0.098***	0.070***	0.084***	0.084***	0.053***
	(0.012)	(0.012)	(0.011)	(0.013)	(0.013)	(0.012)
MTEL x TOC				0.055	0.052	0.059*
				(0.035)	(0.035)	(0.032)
N	38,804	38,804	38,731	36,320	36,320	36,254
School-Grade Group FE		Y			Y	
School-Grade FE			Y			Y

Appendix Table A.5. Licensure Tests and Teacher Performance (First-Time Passing Status)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Value Added						
Pass	0.031***	0.020***	0.018***	0.031***	0.020***	0.017***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
Pass x TOC				0.012	0.001	0.002
				(0.013)	(0.011)	(0.010)
N	2,002,904	2,002,904	2,002,904	1,865,881	1,865,881	1,865,881
Panel B. ELA Value Added						
Pass	0.019***	0.011***	0.007***	0.020***	0.012***	0.007***
	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Pass x TOC				-0.004	-0.007	-0.007
				(0.013)	(0.011)	(0.011)
N	1,880,311	1,880,311	1,880,311	1,726,105	1,726,105	1,726,105
Panel C. Teacher Performance Ratings						
Pass	0.119***	0.117***	0.095***	0.108***	0.107***	0.084***
	(0.006)	(0.006)	(0.005)	(0.007)	(0.007)	(0.006)
Pass x TOC				0.063**	0.057**	0.050**
				(0.026)	(0.026)	(0.022)
N	776,806	776,806	776,169	698,422	698,422	697,843
School-Grade Group FE		Y			Y	
School-Grade FE			Y			Y

Appendix Table A.6. Licensure Tests and Teacher Performance (Alternative Race/Ethnicity Identifiers)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math Value Added						
MTEL	0.024***	0.015***	0.012***	0.024***	0.014***	0.011***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
MTEL x TOC	0.005	0.005	0.006			
	(0.007)	(0.006)	(0.006)			
MTEL x Asian/PI				0.023	0.013	0.012
				(0.014)	(0.012)	(0.011)
MTEL x Black				0.002	0.000	0.002
				(0.014)	(0.012)	(0.012)
MTEL x Hispanic				0.004	-0.004	-0.004
				(0.013)	(0.011)	(0.011)
N	2,002,904	2,002,904	2,002,904	1,865,881	1,865,881	1,865,881
Panel B. ELA Value Added						
MTEL	0.014***	0.007***	0.005**	0.014***	0.007***	0.003*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
MTEL x TOC	-0.001	0.001	0.001			
	(0.008)	(0.007)	(0.007)			
MTEL x Asian/PI				-0.005	-0.006	-0.009
				(0.022)	(0.018)	(0.017)
MTEL x Black				0.011	0.013	0.013
				(0.012)	(0.011)	(0.010)
MTEL x Hispanic				-0.010	-0.007	-0.005
				(0.012)	(0.012)	(0.011)
N	1,880,311	1,880,311	1,880,311	1,726,105	1,726,105	1,726,105
Panel C. Teacher Performance Ratings						
MTEL	0.098***	0.097***	0.073***	0.097***	0.096***	0.072***
	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)

MTEL x TOC	0.030**	0.026*	0.029**			
	(0.014)	(0.015)	(0.012)			
MTEL x Asian/PI				0.060*	0.059*	0.033
				(0.033)	(0.033)	(0.029)
MTEL x Black				0.046	0.043	0.035
				(0.030)	(0.030)	(0.026)
MTEL x Hispanic				-0.019	-0.023	-0.017
				(0.026)	(0.026)	(0.023)
N	776,806	776,806	776,169	698,422	698,422	697,843
School-Grade Group FE		Y			Y	
School-Grade FE			Y			Y

Notes: Regressions of teacher value added on average first-time MTEL scores. Columns 1–3 use teacher race/ethnicity reported by school districts in the administrative data. Columns 4–6 use disaggregated race/ethnicity indicators constructed from self-reported race/ethnicity in the testing data. Covariates in value-added regressions are described in the text. Regressions are weighted by the inverse of the number of observations for each teacher. Standard errors clustered by teacher in parentheses. * $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$.