AMERICAN INSTITUTES FOR RESEARCH®

Analysis of Longitudinal Data in Education Research Program

Massachusetts Educator Preparation and Licensure

Year 1 Report

James Cowan
Dan Goldhaber
Roddy Theobald

MAY 2017

Massachusetts Educator Preparation and Licensure

Year 1 Report

May 2017

James Cowan
Dan Goldhaber
Roddy Theobald

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

Contents

Tables

Page

Figures

Executive Summary

Background

Preparing and licensing teachers is an important leverage point for state education agencies in influencing the quality of the teacher workforce. This is particularly so in Massachusetts, whose diverse teacher preparation landscape includes about 70 educator preparation providers (EPPs) and 2,000 distinct programs within these EPPs at both traditional programs housed in colleges and universities and other providers that offer alternative routes into the profession. This report describes results from the first year of research in a partnership between the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) and the Department of Elementary and Secondary Education (ESE) to study teacher preparation and licensure in Massachusetts.

Specifically, we examine how three outcomes—student achievement on state standardized tests, summative performance ratings from educator evaluation, and teacher attrition—vary among teacher candidates from different licensure pathways, preparation program types, and specific EPPs and programs.[1] These findings are based on a statewide analysis of Massachusetts teachers who completed educator preparation programs or earned a first teaching license in Massachusetts between 2010 and 2014.

It is important to emphasize that there are limitations to this non-experimental research. Although the methods used in this report are designed to separate the effectiveness of teachers from the context in which they work, the results are sensitive to different assumptions (described more fully in the report) about the influence of schools on each of the performance measures. In addition, estimates of program and pathway effects reflect more than just differences in the quality of teacher preparation, because programs and pathways may differ in the academic preparation, innate teaching skills, or prior teaching experiences of their candidates out of state. This summary focuses primarily on results that are consistent across different assumptions, and the limitations of this research are described in more detail in the last section of this summary and in the report.

Key Findings

Key Finding #1: Relative to teachers who receive their teaching license from an undergraduate program, teachers who receive their teaching license from postgraduate or alternative programs tend to receive higher summative performance ratings. Evidence of a relationship between program type and value added is mixed.

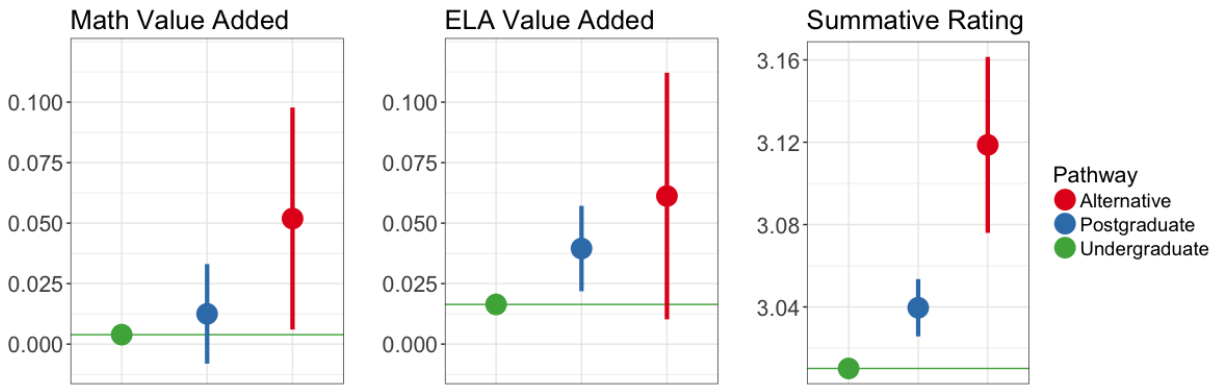
Teachers from postgraduate programs are more effective at raising ELA achievement (but not math achievement) than teachers from undergraduate programs. In Figure ES.1., we plot the

[1] Throughout the report, we follow the conventional terminology and refer to estimated effects on student achievement as "value added." These estimates are derived from statistical models run for this research project and not from the state-calculated student growth percentiles.

expected outcomes on each performance measure for alternative, postgraduate, and undergraduate institutions from our baseline models. The first two panels illustrate findings from the value added models, but these results tend to be sensitive to methodological choices. Although the baseline results shown suggest that teachers from alternative programs have students with higher student achievement gains, this result is not supported by models with more robust controls for student background than the one shown here. Taken as a whole, the results do not provide consistent evidence of differences in value added across program pathways.

The differences across pathways on the summative ratings are more consistent. Teachers who receive their first teaching license from a postgraduate or alternative program tend to receive higher summative performance ratings than teachers who receive their teaching credential through an undergraduate program. We measure performance on each of the standards in the summative ratings by awarding a point for each rating category (unsatisfactory is coded as a 1; exemplary is coded as a 4) and then adjust for school characteristics and district rating standards. Teachers from postgraduate programs earn ratings about 0.03 points higher than those from undergraduate programs, while teachers from alternative programs earn ratings about 0.11 points higher. The difference in summative ratings between undergraduate and postgraduate programs is nearly as large as that between a novice and second year teacher; the difference between undergraduate and alternative programs is similar to that between a novice teacher and a teacher with five years of experience.

Figure ES.1. Program Pathways and Teacher Outcomes



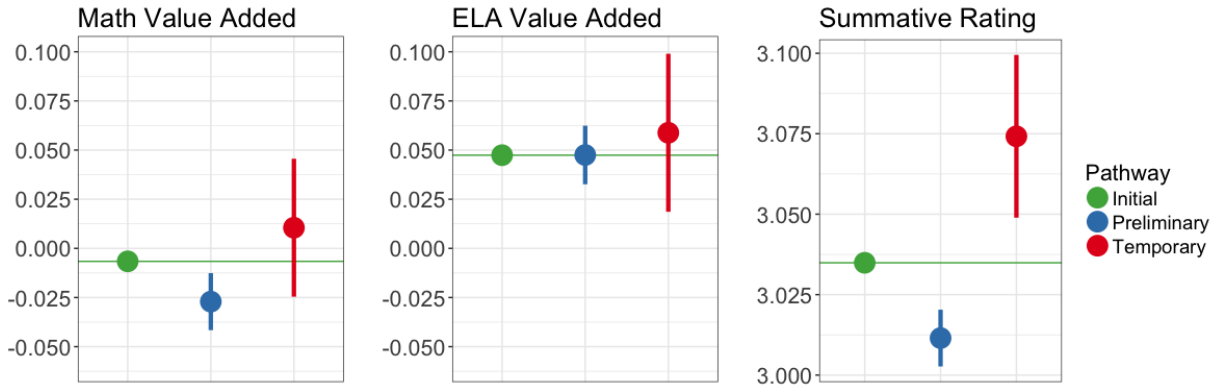
Notes: We plot estimated teacher outcomes for each program pathway at the average values in the dataset. Effects for alternative and postgraduate programs are estimated using baseline models presented in the text. Value added is defined in standard deviations of student test performance; summative performance rating effects are in mean points on the rating scale across the four standard scores; attrition effects are in probability units. The vertical bars depict the 95% confidence interval for the estimated pathway effects relative to undergraduates.

Key Finding #2: Relative to teachers with initial licenses, out-of-state teachers with temporary licenses tend to receive higher summative performance ratings, while those with preliminary licenses that do not require prior program completion tend to receive lower summative performance ratings. Evidence for value added is mixed, but there is some evidence that teachers with preliminary licenses are less effective in math instruction.

We consider three licensure pathways: teachers with a preliminary license (offered to teachers who have passed the state licensure testing requirements but not yet completed an approved EPP), an initial license (offered to teachers who have completed a preparation program and state testing requirements), and a temporary license (offered to teachers with at least 3 years of experience in another state who have not completed state licensure testing requirements). We plot mean outcomes by pathway in Figure ES.2.

As is the case with program pathways, no clear differences in value-added emerge. Teachers who enter with a preliminary license have lower math value added than teachers who enter with an initial license, although the difference is modest (less than one month of student learning in these grades). We do not observe differences in ELA value added. There is also little evidence of any difference in value added between teachers with an initial and temporary license. Differences in effectiveness are again more apparent in the summative ratings data. Teachers with preliminary licenses tend to receive lower summative performance ratings, and teachers with temporary licenses receive higher ratings, than new teachers with initial licenses. The difference in ratings between teachers with preliminary and temporary licenses corresponds to about 0.06 points or the difference between a novice and third-year teacher.

Figure ES.2. Licensure Pathways and Teacher Outcomes

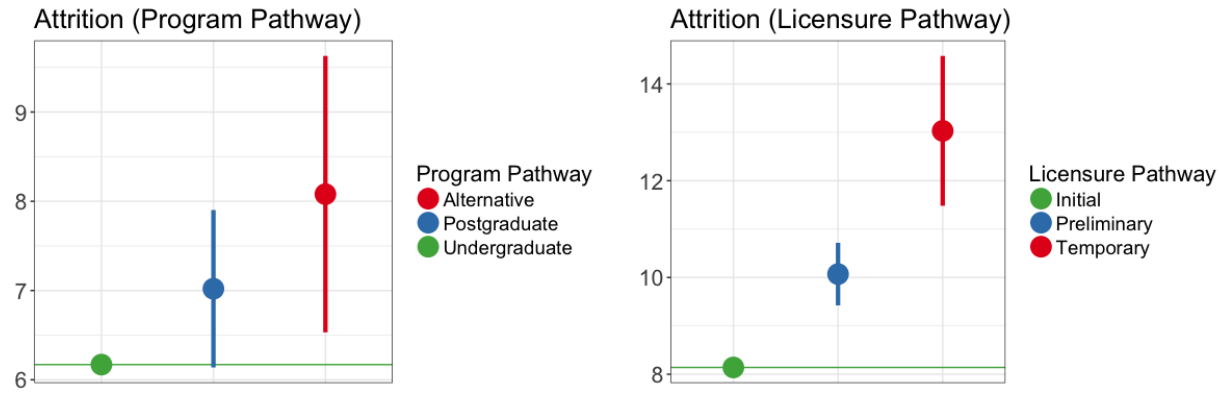


Notes: We plot estimated teacher outcomes for each program pathway at the average values in the dataset. Effects for alternative and postgraduate programs are estimated using baseline models presented in the text. Value added is defined in standard deviations of student test performance; summative performance rating effects are in mean points on the rating scale across the four standard scores; attrition effects are in probability units. The vertical bars depict the 95% confidence interval for the estimated pathway effects relative to initial licenses.

Key Finding #3: Teachers who enter the profession through alternative programs or with non-standard licenses are more likely to leave teaching than teachers from traditional, in-state EPPs.

A consistent finding is that teachers entering the profession with an initial license after completing an approved EPP have the lowest rates of attrition. We illustrate differences in attrition rates by program and license type in Figure ES.3. The average early-career attrition rate for teachers completing in-state programs is about 6% (a probability of 0.06). For example, teachers from alternative programs are about 2 percentage points – or about 33% – more likely to leave teaching in Massachusetts schools in a given year than teachers from undergraduate programs. Similarly, the second panel in Figure ES.3 shows that teachers with preliminary and temporary licenses are also more likely to exit teaching than those with initial licenses. In fact, teachers from outside of Massachusetts are nearly 60% more likely to exit than those with initial licenses from in-state EPPs.

Figure ES.3. Pathways and Teacher Attrition



Notes: We plot estimated teacher attrition for each program and licensure pathway at the average values of teacher covariates in the dataset. Effects for alternative, postgraduate, preliminary, and temporary pathways are estimated using baseline models presented in the text. Attrition effects are in probability units. The vertical bars depict the 95% confidence interval for the estimated pathway effects relative to postgraduate programs or initial licenses.

Key Finding #4: Outcomes for most EPPs in Massachusetts are not statistically distinguishable from the average outcome across EPPs, but the variation explained by individual EPPs in Massachusetts is on the high end of what has been found in other states.

The teacher effectiveness measures for most EPPs in Massachusetts are not statistically distinguishable from the average across Massachusetts EPPs, a finding that is consistent with research in several other states. Although a few EPPs stand out at each extreme, differences in average value added among the EPPs in the middle of the distribution may be driven by random year-to-year fluctuations in the performance of individual teachers or students. In the full report, we discuss alternative methods of estimating differences in EPP outcomes; some program estimates vary significantly across modeling choices. In particular, estimates for EPPs with low enrollment or geographically concentrated placements are especially sensitive to modeling choices.

Despite the fact that most estimated EPP performance measures are statistically insignificant, there are educationally significant differences among programs. Researchers have estimated that typical annual learning gains in the grades we consider are about 0.40 standard deviations per year in math and about 0.30 standard deviations per year in ELA. Given those figures, the estimated achievement differences between the most effective EPPs and the state average correspond to about 5 to 20 weeks of student learning in math and about 9 to 36 weeks of learning in ELA. Overall, the variation in student achievement gains explained by individual EPPs is on the higher end of what has been observed in other states. The EPPs and individual programs jointly explain about 10 to 25% of the variation in teacher value added and about 2% of the variation in summative performance ratings. The institution and program a candidate attended predict future effectiveness about as well as other measures that can be collected during the teacher recruitment process (e.g., endorsements, educational attainment, test scores, and teacher screening tools).

Key Finding #5: Different programs within the same EPP vary substantially in teacher value added.

The individual EPP estimates mask considerable diversity within institutions. In the prior section, we characterize the variation in teacher effectiveness across institutions in Massachusetts. We also conduct a similar analysis to compare average effectiveness for different programs in the same institution. For example, if an EPP prepares candidates in five different licensure areas or prepares candidates at both the graduate and undergraduate levels, we can estimate the average effectiveness measures for each of these groups separately. When we do so, we find that teacher outcomes vary as much from program to program as they do across institutions.

This means that average teacher effectiveness is more variable among programs within an institution than across EPPs.

When the program portion of the variance is large, as it is in Massachusetts, teachers tend to look more like other teachers from their program and less like teachers from other programs in their institution. The empirical importance of individual programs implies that recruitment, preparation, or placement factors that differ among programs within an institution are more consequential than those that are fixed for all programs. Program-specific policies may therefore be an important focal point for improving teacher preparation.

Implications

As noted above, these estimates should be interpreted as reflecting both the effects of teacher training and pre-existing teacher candidate characteristics, rather than causal effects of teacher preparation. Nonetheless, the key findings discussed above have a number of potential implications for policymaking in Massachusetts:

- The variability in teacher outcomes among providers in this study is toward the higher end of comparable analyses in other states, so provider accreditation and support may be a more promising policy lever in Massachusetts than elsewhere.

- Individual programs within EPPs vary substantially in the effectiveness of their graduates. Policymakers may therefore wish to consider the strength of both programs and EPPs when evaluating teacher preparation. Furthermore, to the extent that teacher effectiveness varies across programs within an institution, EPPs may benefit from careful study of the features that distinguish their more effective programs.
- The providers and pathways that produce more effective teachers are not always those whose teachers remain in the profession the longest. In fact, the EPP teacher effectiveness measures are positively correlated with teacher attrition; in other words, the EPPs that graduate more effective teachers also tend to graduate teachers who are more likely to leave the workforce. Nonetheless, the available empirical evidence suggests that the impact of attrition on student outcomes is likely small relative to the direct effects on placing students with more effective teachers. Differences between programs and pathways in terms of teacher attrition are therefore likely less important to student outcomes than differences in their direct effects on student achievement.

The findings in this report provide a framework for considering the relationship between pre-service preparation experiences and teacher effectiveness. However, we only consider a narrow range of outcomes. When reviewing individual EPPs, including multiple indicators of effectiveness may provide policymakers with more reliable judgments about program effectiveness. Using a variety of measures also broadens the range of skills an evaluation system might consider. Emerging research on teacher effectiveness suggests that teaching is multidimensional and that important teaching skills are not fully captured either by test-based measures or by classroom observational tools.

Introduction

The Massachusetts Department of Elementary and Secondary Education (ESE) has identified improving the quality of educator preparation as a key component of its strategic plan. This initiative is consistent with more than a decade of educational research demonstrating that teachers are one of the most important schooling factors affecting student achievement and with emerging evidence that teacher preparation may be an important lever for improving the teaching workforce. The National Center for the Analysis of Longitudinal Data in Education Research (CALDER) at American Institutes for Research has partnered with ESE to study the relationship between educator preparation in Massachusetts and student and teacher outcomes in the state. In this report, we describe results from the first year of research, which focuses on providing an overview of the educator preparation system in Massachusetts. In particular, we provide descriptive evidence of how teacher candidate outcomes vary across licensure pathways, preparation program types, and specific educator preparation institutions and programs in the state.[2]

Research Questions

To define the specific questions for the first year of this collaboration, CALDER staff reviewed the ESE strategic plan, proposed several potential topics drawn from the broader literature on teacher preparation, and then worked with ESE to refine the proposed research questions to best support ESE priorities. CALDER and ESE jointly identified three key workforce outcomes that guide this research: student achievement on annual standardized tests; teacher summative performance ratings; and teacher retention in the public school system.

During the first year of this research grant, we have taken a broad view of the preparation and licensure landscape in Massachusetts and assembled descriptive evidence on how these key workforce outcomes vary with the nature of teachers' preservice preparation experiences.[3] Specifically, we have investigated three primary research questions:

1. What is the variation in student achievement gains associated with different teacher licensure pathways, preparation program types, and specific educator preparation institutions and programs?

2. What is the variation in teacher evaluation results associated with different teacher licensure pathways, preparation program types, and specific educator preparation institutions and programs?

[2] Throughout this report, we refer to an institution authorized to recommend graduates for licensure, such as a university or organization offering alternative licensure programs, as a *preparation provider, institution,* or *organization* and to a specific course of study within an institution leading to a license in a teaching field, such as *Elementary Teacher, Grades 1–6*, as a *program*. An institution may therefore include multiple programs.

[3] We refer to the results presented in this report as descriptive because, as discussed in the next section, all results in this report reflect a combination of the selection and training effects of licensure pathways, educator preparation institutions, and educator preparation programs.

3. What is the variation in teacher retention associated with different teacher licensure pathways, preparation program types, and specific educator preparation institutions and programs?

Each of these research questions relates three different descriptors of teachers' preservice preparation to one of the key outcomes identified by CALDER and ESE. The first descriptor is a teacher's route into the profession, which we classify using the type of the license a teacher first earns in Massachusetts. For the remaining analysis, we limit our focus to teachers completing their teacher preparation in Massachusetts. The second descriptor is the type of preparation program attended—undergraduate, postgraduate, and alternative—and the third is a teacher's specific education institution and program. Although we estimate and report individual contrasts in educator effectiveness and retention rates between particular institutions as part of research question #3, the broader goal of this research is to quantify the variability in educator outcomes across and within institutions and programs in Massachusetts. We therefore refer to individual institutions using anonymous identifiers throughout this report.

Background

Teacher Effectiveness and Student Learning

More than a decade of empirical research using statewide databases has consistently found substantial variation in test score gains between students with different teachers in a variety of states, subjects, and grade levels. These differences in educator effectiveness comprise one of the most variable schooling inputs in the research literature (Goldhaber et al., 1999; Rivkin et al., 2005). Differences in teacher quality explain about 1–5% of the overall variance in student achievement (e.g., Aaronson et al., 2007; Chetty et al., 2014a; Hanushek & Rivkin, 2010). Although this proportion may seem small, the expected impact of a one standard deviation increase in teacher quality is roughly equivalent to the effect of reducing class size by about 10 students (Rivkin et al., 2005). More recent research has also linked teachers to a variety of other important student outcomes, including future performance on academic tests as well as discipline, grades, educational attainment, and earnings (Chetty et al., 2014b; Gershenson, 2016; Kane et al., 2013; Jackson, 2016; Kinsler, 2012; Papay, 2011).

Unfortunately, a large research base has found inconsistent or weak connections between a teacher's impact on student test scores and traditional teacher credentials, such as licensure status and degree level (e.g., Goldhaber, 2002; Hanushek, 1986; Rivkin et al., 2005). Likewise, in-service interventions designed to improve teacher effectiveness—such as professional development (e.g., Garet et al., 2016; Hill & Ball, 2004; Jacob & Lefgren, 2004) and pay for performance (e.g., Glazerman & Seifullah, 2010; Goldhaber & Walch, 2012; Springer et al., 2011)—have also been found to have minimal impacts on student outcomes. But while the studies cited above—along with the vast majority of research on teacher quality—focus on teachers who are *already in the profession*, much of a state's investment in teacher workforce development occurs *before* teachers enter the workforce. This helps to motivate the focus on educator preparation in this research project.

Educator Preparation in Massachusetts

Massachusetts is an interesting setting to explore the connections between teachers' preservice experiences and their classroom effectiveness and retention given the number and diversity of institutions preparing teachers in the state. There are about 70 different institutions involved in teacher education, including 16 alternative providers that operate outside traditional higher education settings. These include several high-profile providers such as the Boston Teacher Residency and Teach for America. Altogether, institutions in Massachusetts provide more than 2,000 teacher training programs (U.S. Department of Education Office of Postsecondary Education, 2014). This environment encompasses a greater number of individual programs and more heterogeneity in settings than those considered in most existing studies of traditional teacher preparation programs.[4] Although all educator preparation programs in Massachusetts are designed to meet the same program approval standards, there may be more variation in program design and curricular requirements across licensure pathways or program types than among programs within the same pathway. Therefore, before investigating differences in outcomes between institutions, we first compare teachers who enter through each of these alternative pathways to those who complete a traditional educator preparation program.

Beyond the range of educator preparation program types, Massachusetts also offers two additional licenses for teachers who have not completed a state-approved educator preparation program. The *temporary* license is valid for 1 year for teachers who have an out-of-state license and more than 3 years of teaching experience but who have not completed the Massachusetts Tests for Educator Licensure (MTEL) requirements. The *preliminary* license, which is valid for 5 years, permits teachers who have completed the MTEL requirements but not an educator preparation program to teach in the state. Teachers entering the profession with different classes of teaching license have therefore likely had substantially different prior teacher education or teaching experience.

Educator Preparation and Teacher Effectiveness

The role of teacher licensing requirements in ensuring the overall quality of teaching has long been an issue of debate.[5] Much of the recent evidence on the effectiveness of licensure requirements comes from comparisons of fully certified teachers to those with provisional or temporary credentials. The results from this literature are mixed and likely partially depend on policy context, including licensure requirements and the recruiting practices of alternative certification programs. Studies of groups of alternative entry or lateral entry teachers in other states have produced inconsistent results (Bastian & Henry, 2015; Boyd et al., 2006; Clotfelter et al., 2006, 2010; Goldhaber & Brewer, 2000; Kane et al., 2008; Sass, 2015), although a few random assignment studies have found these groups not to be statistically significantly less effective than traditionally certified teachers (Clark et al., 2013; Constantine et al., 2009). Although the statistical significance—and sometimes direction—of these results differ, the

[4] Analyses from Florida (Mihaly et al., 2013) and Tennessee (Ronfeldt & Campbell, 2016) are similar in terms of the number of individual institutions they include; however, we observe far more individual educator preparation programs in Massachusetts.

[5] For instance, see National Commission on Teaching and America's Future (1996) and Ballou and Podgursky (1998). More recent empirical analyses of state licensing requirements include Angrist and Guryan (2008) and Larsen (2015).

differences in effectiveness are typically, although not uniformly, less than those between first-year and third-year teachers. There is also some evidence that initial deficits in the effectiveness of alternatively certified teachers may diminish as they gain teaching experience (Boyd et al., 2006; Papay et al., 2012). Although the policy context appears to matter for the empirical results on licensure, the main point of commonality among these studies is that the majority of the variation in teacher quality is within, rather than between, categories of teacher preparation.

In addition to general analyses of the effectiveness teachers entering the profession through alternative pathways, there have been a number of studies of particular alternative programs. As is the case with the alternative pathway literature generally, there is considerable heterogeneity in the effectiveness of particular programs (e.g., Boyd et al., 2006; Sass, 2015). We briefly mention a few studies that include programs operating in Massachusetts. There is a considerable amount of literature on Teach for America, which recruits teachers from selective colleges to teach in high-poverty schools. Evidence from random assignment experiments (Clark et al., 2013; Glazerman et al., 2006) and observational studies (Boyd et al., 2006; Kane et al., 2008; Xu et al., 2011) suggests that these teachers are more effective at math instruction than teachers in their schools with regular certification. The evidence on reading or English language arts, however, is more mixed (e.g., Boyd et al., 2006; Glazerman et al., 2006). Papay et al. (2012) studied early cohorts of the Boston Teacher Residency and found that they are initially less effective in math subjects only, but this deficit shrinks over time.

Another line of research has examined the variability of traditional educator preparation institutions and programs in several states. This research suggests that there are differences in average effectiveness across institutions (Boyd et al., 2009; Goldhaber et al., 2013; Koedel et al., 2015b; Mihaly et al., 2013) and programs (Henry et al., 2014). However, the magnitude of variation differs across states and by research methodology, and the substantive importance of these differences remains unclear. On the higher end, Boyd et al. (2009) estimated that the preparation institution explains about 15% of the variation in teacher value added in New York City. At the other extreme, Koedel et al. (2015b) found that preparation programs in Missouri explain only 0% to 3% of the variation in teacher effectiveness. More recently, researchers have begun linking other in-service performance measures to teacher educator preparation institutions and programs. Using data from Tennessee, Ronfeldt and Campbell (2016) found that about 3% to 4% of the variability on the state's classroom observation rubric is attributable to preparation institutions. Bastian et al. (2015) also established a link between teacher evaluation results and educator preparation programs.

Educator Preparation and Retention

The final outcome we consider in this study is teacher retention in the Massachusetts public school system. Teachers leave the profession at a higher rate during the first few years in the classroom. According to the most recent Schools and Staffing Survey, about 7% of novice teachers (those with 1 to 3 years of experience) left the profession between 2012 and 2013 (Goldring et al., 2014). Teacher retention has a number of direct and indirect costs to districts. Barnes et al. (2007) estimate that the costs of replacing departing teachers may be close to $10,000 in some districts. Beyond the costs to districts in terms of recruiting, hiring, and training new teachers, teacher turnover may harm student achievement through at least two channels. First, departing teachers are frequently replaced by novice teachers, who tend to be less effective

than teachers with even a few years of experience (Clotfelter et al., 2007; Rockoff, 2004). Second, turnover may disrupt collaborative relationships or otherwise harm school climate (Ronfeldt et al., 2013). But these mechanisms affect students in other classrooms as well, which means that the full effect of teacher turnover will not be reflected in the teacher effectiveness measures for a particular institution. We therefore consider teacher turnover as a separate outcome.

There is substantial empirical evidence that preservice experiences correlate with teacher retention. Out-of-state teachers in North Carolina (Bastian & Henry, 2015) and Washington state (Goldhaber & Cowan, 2014) have higher attrition rates than teachers from in-state institutions. Teachers with alternative teaching credentials also appear to have higher attrition rates (Bastian & Henry, 2015; Boyd et al., 2006; Kane et al., 2008; Redding & Smith, 2016). This appears to be particularly true for teachers who enter the profession through Teach for America, which has an initial 2-year commitment (Donaldson & Johnson, 2011; Hansen et al., 2016; Kane et al., 2008). Researchers have also begun to link more specific elements of preservice preparation to teacher retention. For example, "high-quality" student teaching placements (as measured by student teacher perceptions or schools with low staff turnover) and methods coursework appear to improve retention once teachers enter the classroom (Ronfeldt, 2012; Ronfeldt et al., 2014; Goldhaber et al., 2016).[6] Retention rates also vary meaningfully across preparation institutions (Goldhaber & Cowan, 2014). Thus, not only does attrition vary with the amount of preservice preparation, but it also varies with experiences that may differ across traditional preparation programs. Taken together, these findings suggest that teacher retention is likely to vary across program and licensure pathways, but also across individual institutions and programs.

Selection and Training Effects

Educator preparation programs and institutions serve multiple roles in the teacher licensing infrastructure. They act as gatekeepers into the profession by selecting candidates for admission and setting internal standards for program completion. They design curricula and provide teachers with their first professional training as educators. They may also operate as professional networks and influence whether or where teacher candidates obtain positions in the public school workforce (Goldhaber et al., 2014; Krieg et al., 2016). The capacity of different institutions and programs to carry out each of these roles likely influences the workforce outcomes of their graduates. Short of randomly assigning teacher candidates to attend particular programs, disentangling the effects of the selectivity of an institution from the effect of its training, curriculum, or other practices is not straightforward.

Our estimates of teacher outcomes by program or pathway therefore likely represent the contributions of several factors. Some of these, which we term *training effects*, are related to the quality of instruction candidates receive in their programs. The recent research linking preservice experiences to workforce outcomes suggests that these practices vary among teacher candidates and that these experiences meaningfully affect classroom practice and retention. The effects of faculty, curriculum, and student internships will be reflected in our estimates of institution

[6] The effectiveness of educator preparation practices remains an important ongoing field of research, but there is evidence that conducting student teaching in schools with characteristics that proxy for collegiality and school climate improves teaching practice (Ronfeldt, 2012).

outcomes. But institutions may also differ in their admissions requirements or in the applicants they attract. Undergraduate programs at highly selective institutions may admit candidates with stronger academic backgrounds. Admissions to postgraduate programs may differentially weight prior teaching experiences or other signals of a candidate's commitment to the teaching profession. These *selection effects* are likely also related to teacher effectiveness and retention. Some institutions may admit a pool of candidates that would have strong outcomes regardless of the quality of training they receive. Thus, differences in teacher outcomes across institutions may also reflect innate teaching skills or other prior experiences of candidates before they enter their programs.

Beyond the influence of selectivity and training, there are at least two sources of non-random sorting into Massachusetts public schools that might influence our results. First, preparation programs differ in the likelihood that their candidates will obtain teaching positions in Massachusetts.[7] The administrative data we use in this study only accounts for the outcomes of those teachers who eventually obtain jobs in the public school system. For programs that disproportionately place students in private schools, out-of-state schools, or other professions, our estimates may not be representative of their graduates as a whole. For instance, a program with a poor track record of placing candidates may only obtain teaching positions for its most effective graduates. In this case, the average effectiveness of the teachers we observe is likely higher than among completers in the program as a whole. However, given the role of the programs in the teacher licensure system, we are interested in predicting effectiveness of those working in public schools. Second, programs may differ in the types of schools in which their candidates work. For example, some programs disproportionately place their graduates in charter schools, schools in disadvantaged districts, or exam schools. Failing to appropriately adjust for school context may lead the estimated institution effects to conflate the effectiveness of teachers with the effectiveness of the schools in which they teach. We employ several empirical methods, described more fully in the methods section, to account for this latter source of selection.

The conceptualization of institution effects as incorporating the implications of selection and training follows much of the recent literature on the analysis of educator preparation programs using administrative data (e.g., Boyd et al., 2009; Goldhaber et al., 2013b; Koedel et al., 2015b). Our underlying objective is to predict the effectiveness or retention of teachers from a particular institution or pathway taking its selectivity, placement patterns, and training as given. This is consistent with appreciating the multiple responsibilities of preparation providers as gatekeepers and educators. Therefore, these measures do not solely reflect the contributions of training at a particular educator preparation institution or program or within a particular licensure pathway or program type. That is, they likely do not represent the "causal effects" on teachers of attending a specific program. In particular, we would not expect that our estimates would predict the results of a hypothetical experiment that randomly assigns potential teacher candidates to licensure pathways, program types, and specific institutions and programs.

We instead focus on adjusting for posttraining experiences, such as educational resources, student background, or school quality, that may be correlated with teachers' preservice

[7] In our sample, the placement rate is 67%. However, this varies considerably across institutions. Among the largest institutions we consider (those that contribute at least 50 teachers to each sample), placement rates range from 54% to 84%.

experiences but have no clear causal relationship with them. These measures are of use to administrators and policy makers who wish to predict teacher workforce outcomes for candidates from different licensure pathways, program types, and specific institutions and programs. Understanding whether teachers who enter the state's teaching workforce through different licensure pathways, program types, and specific institutions and programs perform especially well or poorly relative to other teachers in the state—even if we cannot identify specific causal mechanisms—might help policy makers to regulate teacher licensure in the state and identify institutions or programs that may technical assistance. School districts may also benefit from knowing which pathways are associated with teacher effectiveness and turnover. These analyses may also help generate hypotheses about the attributes of effective educator preparation.

Data and Sample

In this study, we focus on cohorts of teachers who completed educator preparation programs or earned a first teaching license in Massachusetts between 2010 and 2014. A number of factors influenced our choice of cohorts. The most influential of these was the availability of data linking students and teachers. These data are necessary to perform the analyses of student achievement in this study. The Massachusetts databases began linking students and teachers in 2011, and we can therefore observe all student outcomes in the classrooms of teachers who earned a first teaching license in Massachusetts between 2010 and 2014. The use of these recent cohorts also ensures that the analyses reflect the recent functioning of programs, although it is possible that program features have changed since the Massachusetts Board of Elementary and Secondary Education approved new program approval standards in 2012. The use of additional cohorts may improve the precision of our estimates, but only at the cost of considering increasingly distant graduates. The window considered in this study is also likely to be more similar to those used for program accountability and evaluation purposes.

Table 1 contains the total count of teachers included in this study: the first two rows identifies teachers who are included in the value-added models (VAMs) in math and English language arts (ELA); the third row identifies teachers included in the analysis of teacher summative evaluations; and the last row identifies teachers included in the attrition analysis. For each row, the first column identifies teachers for whom we observe their preparation institution and program, while the second column identifies teachers for whom we observe their licensure pathway.

Table 1. Sample Sizes by Analytic Sample

Analysis Type	Number of Teachers	
	Preparation Program Analysis	Licensure Analysis
VAM (Math)	2,377	4,748
VAM (ELA)	2,353	4,500
Summative Evaluation	8,880	17,431
Attrition	7,640	13,991

Note: Table presents counts of teachers in each of the analysis samples by pathway type. The licensure analysis sample includes all teachers whose first Massachusetts teaching license was earned between 2010 and 2014 and was an initial, preliminary, or temporary license. The preparation program sample includes all teachers who

completed an educator preparation program in Massachusetts and earned their first initial license of the same type within 6 months of completion. VAM = value-added model.

In the following sections, we provide an overview of the construction of the licensure pathway and program completer samples. The program completer sample is a subset of the licensure sample and includes teachers we can verifiably link to program completion records in Massachusetts. We explain the procedures for matching completers to licensure records and the sample restrictions below. We then describe the data sets we use to study each of the research questions. These data sets correspond to the four rows in Table 1 and each relies on a different subset of teachers. For instance, attrition data are available for all teachers in four of the five study cohorts, but value-added data are available only for teachers in tested grades and subjects.

Licensure Pathway Sample

We identify teachers' licensure pathway based on the first entry-level teaching credential they earned in Massachusetts. Our analysis of licensure pathways uses data on all teachers earning their first teaching license in Massachusetts between 2010 and 2014. We consider three types of licenses in this study. The most common license, the *initial license*, is awarded to teachers who have completed all of the requirements for a teaching license in Massachusetts. Depending on the sample, teachers with initial licenses account for about 57% of the teachers in this study. Teacher candidates who have passed the Massachusetts Tests for Educator Licensure (MTEL) but have not completed an educator preparation program are eligible for the *preliminary license*. The preliminary license is valid for 5 years and cannot be renewed. In order to remain in the teaching profession, teachers must complete an educator preparation program and graduate to an initial license within the 5-year period. The preliminary license is the most common among teachers who enter the workforce through alternative licensure pathways in Massachusetts. These teachers comprise about 40% of the licensure sample. Finally, teachers who have not completed the MTEL requirements but have completed an educator preparation program are eligible for the *temporary license*. The temporary license is valid for 1 year with no renewal options and is appropriate for experienced out-of-state teachers who need additional time to complete the licensure test requirements. Only about 3-4% of teachers in our sample hold the temporary license.

Table 2. Average School Demographics by Licensure Pathway

License Type	Percent Hispanic	Percent Black	Percent Asian	Percent Free or Reduced Price Lunch	Percent English Language Learner	Percent Special Education
Initial	19.9	10.4	6.0	44.8	10.3	17.5
Preliminary	23.6	11.9	5.2	50.5	10.3	18.9
Temporary	17.7	10.2	6.3	39.6	9.3	17.9
State Average	21.3	11.0	5.7	47.0	10.3	18.1

Note: Table contains average school characteristics for teachers in the teacher attrition sample. School aggregates were calculated by the authors using student-level data. Observations are at the teacher-year level.

We present mean school characteristics by licensure pathway for the licensure sample in **Table 2**. The summary statistics are based on the attrition sample and are averaged over all years a teacher appears in the dataset. Teachers with preliminary licenses, who have not yet completed their educator preparation, tend to teach in schools with higher proportions of subsidized lunch and Hispanic and Black students. On average, teachers with preliminary licenses teach in schools where about 50% of students qualify for subsidized lunch. Teachers with initial licenses teach in schools with average subsidized lunch participation rates of about 45%. Teachers with temporary licenses tend to teach in less disadvantaged schools: the average teacher works in a school where only 40% of students qualify for subsidized lunches. They also work in schools with fewer Hispanic students and English language learners, although in other respects, the schools appear similar to teachers with initial licenses. This pattern is consistent with analyses of the differences by student socioeconomic status in the assignment patterns of highly qualified teachers, which suggests that poor and minority students are more likely to attend classes with less qualified teachers (Clotfelter, Ladd, Vigdor & Wheeler, 2006; Goldhaber et al., 2015).

Preparation Program Completer Sample

We next construct a sample that links educators to their preparation institutions and programs using a database that tracks program completers in Massachusetts. We consider programs that lead to academic teaching licenses in Massachusetts. Given the focus on classroom teachers completing their first preparation program in Massachusetts, we exclude administrator, professional support, and specialist programs as well as endorsement and apprenticeship programs. Within this set of programs, we classify institutions as either *traditional* or *alternative* using a list of alternative programs provided by Massachusetts ESE. The alternative programs in this study serve students who have previously earned a bachelor degree. They are hosted outside of traditional institutes of higher education and tend to introduce classroom teaching earlier in the course of preparation. Although they are based outside colleges and universities, alternative programs meet Massachusetts state curriculum requirements for educator preparation. Teachers from these programs comprise about 5-7% of the teachers in our data.

We additionally classify traditional institutions into two groups: *undergraduate* programs and *postgraduate* programs. Undergraduate programs culminate in a bachelor degree in education and generally enroll students completing their first postsecondary degree. Postgraduate programs serve students who have already earned a bachelor degree, often in another field, and award teacher candidates a post-baccaluareate degree. Teachers often complete postgraduate programs after earning an initial teaching license as districts typically award additional pay for a graduate degree. Because we focus on teachers earning their first teaching credential in this sample, we exclude midcareer teachers from our analysis. As a result, the postgraduate teachers we consider typically do not have much prior teaching experience.

We further subclassify preparation programs by the field (e.g., biology or mathematics), level (e.g., Grades 1–6 or 8–12), and academic level (e.g., baccalaureate or postbaccalaureate) identified in the program completion files. Of the 68 providers in Massachusetts for which we could match program completers to new teaching licensees, our analytic samples include graduates from 51 institutions and 904 individual programs.[8] Although we observe a large

[8] We provide a complete list of the number of completers in each institution and sample in Appendix A.

number of institutions and programs, we observe only a few teachers for many of the institutions considered. For instance, 20 of the 37 institutions contributing teachers to the math value-added sample have 50 or fewer completers. As a consequence, many of the institution indicators are imprecisely estimated. This is a general problem with the estimation of the effectiveness of educator preparation programs (Lincove et al., 2014). In order to avoid reporting estimates that rely on only a very small number of teachers, we omit institutions with fewer than 15 completers.

We define teachers' educator preparation institution and program using the first recorded completion in Massachusetts. Teacher candidates may complete several programs in Massachusetts over the course of their teaching career. For instance, teachers may complete both a baccalaureate program for their initial license and then later complete a postgraduate or academic specialist program. To ensure that we limited our analysis to first-time program completers, we restricted the sample to teacher candidates who qualified for an initial license that matched their program within 180 days of their program completion. This restriction omits teachers whose licensure record indicates they previously completed a program, even if that completion is not recorded within the timespan covered by the Massachusetts administrative data.

We also limited the sample to first-time completers because we do not observe a direct measure of teacher experience. The lack of a teacher experience measure is problematic because research has consistently demonstrated large returns to teaching experience during the early portion of teachers' careers (Clotfelter et al., 2007, 2010; Harris & Sass, 2011; Rockoff, 2004).[9] It is therefore difficult to identify novice teachers in the data with complete certainty. By focusing on teachers who earned an initial license near their program completion date, this restriction helps ensure that we do not mistakenly assign the effects of prior teaching experience that is not recorded in our data to teachers' preparation programs. In addition, focusing on program completers provides a common sample inclusion criterion that applies to both traditional and alternative programs. Candidates from alternative programs may appear as classroom teachers prior to the completion of their educator preparation program.[10] Because of our focus on program graduates, we choose to focus on completers only and test the sensitivity of our results to how we define teaching experience.

In this study, we rely on three proxies for teaching experience. We first measure the number of years since a teacher completed his or her Massachusetts program leading to initial licensure. We refer to this as the *potential experience* measure as it indicates the number of years a teacher may have taught following program completion. This is the primary measure of experience we use in

[9] One problem common to other state databases is that we do not observe experience in private or out-of-state schools. Although these data limitations are shared with most other state administrative data sets, the Massachusetts data do not provide a measure of overall teaching experience. For instance, researchers sometimes rely on a teacher's location on the salary schedule as a proxy for prior teaching experience. This is not available in the Massachusetts data.

[10] Our research objective is to understand differences in teaching effectiveness and retention for educators completing different programs in Massachusetts; for example, we only consider teachers in alternative programs *after* they complete their program. This is necessarily a different research question than the effectiveness of alternatively certified teachers who have not yet completed a course of study. For results on the effectiveness of teachers from alternative routes during the early portion of their career, we refer readers to the studies by Clark et al. (2013), Decker et al. (2004), Glazerman et al. (2006), Hansen and Sass (2015), and Xu et al. (2011) on Teach for America and to research by Papay et al. (2012) on Boston Teacher Residency.

our analysis of educator preparation programs. Our second measure of experience, which we refer to as *observed experience*, indicates the number of years in a teaching position in Massachusetts public schools following licensure or completion of an educator preparation program. We use this measure in our attrition analysis. Finally, we measure *licensed experience* as the number of years a teacher has held a valid teaching license in Massachusetts. For teachers in standard educator preparation programs, the licensed experience measure will typically equal the potential experience measure, although some teachers complete traditional programs after earning a preliminary license. Teachers who complete alternative programs typically earn a preliminary license and work as a teacher of record prior to program completion. For these teachers, licensed experience will generally exceed potential experience by 1–2 years. We use this measure primarily in our analysis of licensure pathways.

Student Achievement Data

In this study, we use student achievement data from the 2011–2015 academic years. For these school years, scheduling data in EPIMS and SIMS provide a link between students and teachers. Using these data, we can connect teachers' preparation data to the academic performance of their students. Similar data sets have supported most of the contemporary research on teacher quality, including analyses of educator preparation programs in Florida, Missouri, New York City, North Carolina, Tennessee, and Washington. To study the association between preparation and student achievement, we draw on student performance on the MCAS and PARCC assessments. Given the use of both current and prior-year student test scores, our analysis focuses on math and ELA test results for teachers in Grades 4–8 and 10 during the 2011–2015 school years. Importantly, the use of test scores as an outcome measure limits the VAM samples only to teachers in these tested grades and subjects, which may not be representative of the teaching profession as a whole.[11]

We match students to teachers using records from the Student Course Schedule (SCS) and EPIMS data collections. These files contain a record for each course taken by a student or taught by a teacher. The course and section codes listed in these files enable us to link students to their classroom teachers. Our sample includes classrooms identified as math, ELA, or self-contained in Grades 4–8 and 10. We exclude English as a second language classrooms and supplemental and developmental classes. In order to ensure that included classrooms represent authentic student-teacher links, we also exclude students enrolled in two separate math or ELA classrooms. Although this restriction drops students who may be receiving additional math or ELA instruction for remedial purposes, we adopt this more conservative approach to minimize incorrect student-teacher matches and ensure that a student's learning gains in that subject are attributable to the teacher identified in the course schedule data.[12] Finally, we exclude students linked to multiple teachers in each subject unless the course is identified as a team teaching arrangement.

[11] Among teachers who complete a program and who we ever observe working in Massachusetts public schools in some capacity, we match 18.8% to a valid math classroom and 18.6% to a valid ELA classroom.

[12] In addition, there has been limited research on methods for estimating value added for teachers working primarily with special populations of students. Loeb et al. (2014), who studied English language learners, is one exception.

We present summary statistics for the samples of students for whom we can estimate value-added models in math or ELA (the math and ELA VAM samples) in **Table 3**. We standardize the test scores within the full tested sample before limiting the data to recent completers. The negative means in the math columns therefore indicate that our sample has slightly lower baseline achievement than the state as a whole, while the positive means on test scores in the ELA columns indicate that this sample has slightly higher baseline achievement. The average teacher in the sample is about 1.3 years removed from completion of an educator preparation program and 1.8 years removed from first licensure. Differences between the two experience measures are driven by teachers earning preliminary licenses before completing a teacher preparation program. In both samples, graduates of traditional preparation programs are the most common. In the completer sample, 65% of teachers graduated from a postgraduate program and 5–7% graduated from an alternative program. In the licensure sample, 53% of math teachers and 65% of ELA teachers entered with an initial license. The preliminary license is common, and it is more common in math (43%) than ELA (32%). Temporary licenses only account for 3–4% of our sample.

Table 3. Summary Statistics for VAM Samples

Program Completer Sample			Licensure Sample		
	Math	ELA		Math	ELA
Posttest	-0.01	0.00	Posttest	-0.04	0.012
Prior Math	-0.01	0.01	Prior Math	-0.03	0.013
Prior ELA	-0.02	0.01	Prior ELA	-0.04	0.019
Pct. Hispanic	16.5	16.9	Pct. Hispanic	18.1	16.6
Pct. Black	9.3	9.0	Pct. Black	9.8	9.0
Pct. Asian	5.5	5.7	Pct. Asian	5.7	5.6
Pct. Free Lunch	35.3	35.3	Pct. Free Lunch	37.0	34.2
Pct. Reduced Lunch	5.8	5.7	Pct. Reduced Lunch	5.9	5.6
Pct. Limited English Proficient	5.3	4.7	Pct. Limited English Proficient	5.9	4.9
Pct. Special Education	17.6	16.5	Pct. Special Education	17.3	17.1
Potential Experience	1.30	1.27	Potential Experience	1.31	1.25
Licensed Experience	1.84	1.88	Licensed Experience	1.81	1.82
Pct. Alternative Prog	7.2	5.2	Pct. Preliminary License	43.4	32.2
Pct. Post-BA Prog	64.6	65.3	Pct. Temporary License	3.6	2.9
N	141,557	128,009	N	282,083	246,971

Note: Table presents summary statistics for VAM samples. The samples include completers from the 2010–2014 cohorts and student achievement data for the 2011–2015 school years. We exclude institutions with fewer than 15 completers. Test scores are standardized by grade and year in the full sample; the mean score statewide is therefore 0 with a standard deviation of 1. Abbreviations: Prog = program, ELA = English Language Arts, BA = Bachelor's. Standard deviations in parentheses.

Summative Performance Data

As an additional measure of teacher quality, we use data on teachers' summative performance assessments under the new Massachusetts educator evaluation framework. Massachusetts began implementation of evaluations aligned to the Professional Standards for Teaching (PST) for a select group of districts in the 2012–2013 school year and statewide in 2013–2014. The PST include four individual standards covering different areas of teaching practice:[13]

- *Curriculum, planning, and assessment* covers content and pedagogical knowledge, lesson planning, and the use and analysis of assessment data.
- *Teaching all students* assesses the classroom environment, student work and engagement, and appreciation for diverse student backgrounds and learning needs.
- *Family and community engagement* includes indicators for the quality of communication with parents and families and their engagement in their children's learning.
- *Professional culture* covers teachers' professional development and contributions to school leadership. These standards form the basis of teachers' summative performance ratings.

The summative performance assessments are available for a broad set of classroom teachers in Massachusetts (as noted above, the analysis of student test score results necessarily limits the sample of teachers to those in grades and subjects that have been tested).[14] Although not specifically based on the Massachusetts standards, empirical evidence from several sources indicates that individual indicators of the Massachusetts educator evaluation system are related to other measures of teacher effectiveness. For example, classroom observations of teacher practice have been shown to predict student achievement gains and students' reports of classroom environment (Blazar, 2015; Grossman et al., 2013; Kane et al., 2011, 2013). Similarly, evaluations by administrators or mentors can predict test-based measures of effectiveness (Harris & Sass, 2014; Jacob & Lefgren, 2008). Finally, assessments like those offered by the National Board for Professional Teaching Standards, which include an evaluation of educators' ability to assess student needs and tailor instruction appropriately, tend to identify teachers who are more effective at raising student test scores (Cantrell et al., 2008; Cowan & Goldhaber, 2015).

The summative evaluation of teachers is one element of the state's broader evaluation process. Evaluation follows a five-step cycle with a timeline that depends on a teacher's career stage.[15] The cycle begins with a self-assessment by the teacher and the development of a professional growth plan. During the implementation of the growth plan, teachers receive feedback through a formative assessment process. Finally, the cycle concludes with a summative evaluation of teaching practice. Teachers receive an evaluation for each of the four standards and an overall

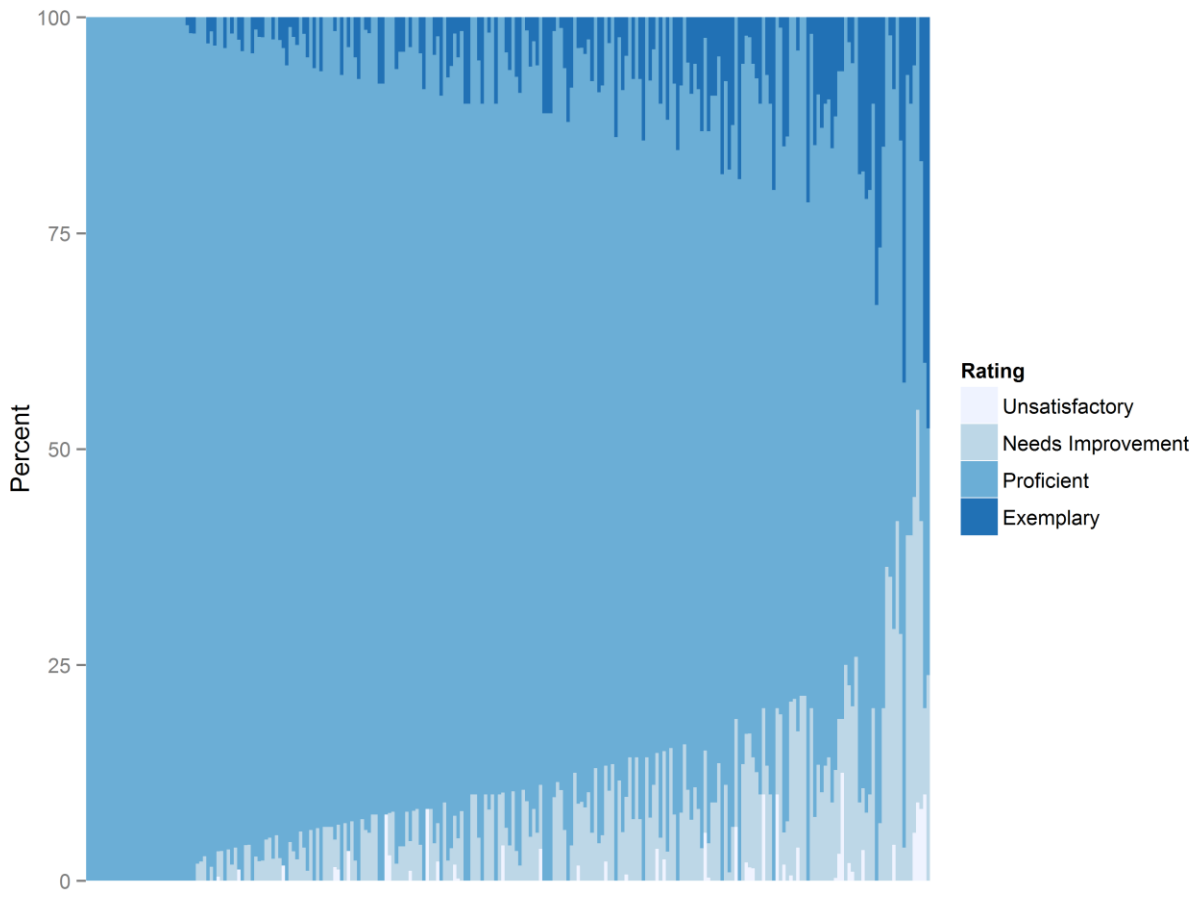
[13] This coverage of the PST follows Massachusetts Department of Elementary and Secondary Education (2015a, 2015e).
[14] Depending on their professional status, teachers in Massachusetts do not necessarily receive summative evaluations in each school year. However, evaluation is more frequent for the sample of novice teachers we study.
[15] This coverage of the educator evaluation framework follows Massachusetts Department of Elementary and Secondary Education (2015b, 2015c, 2015d).

summative performance rating. Each of the standards is rated on a four-point rating scale: *unsatisfactory*, *needs improvement*, *proficient*, or *exemplary*. The state requires that teachers earning a proficient rating must receive at least a rating of proficient on both the curriculum, planning, and assessment and on the teaching all student standards. Beyond this requirement, the evaluation framework preserves an important role for local evaluators in determining how a teachers' performance informs the final rating.

Figure 1 illustrates the variation in scores on the summative performance rating by school district in Massachusetts. We plot the distribution of ratings for teachers in our sample using data from the 2013–2014 and 2014–2015 school years. Each vertical bar represents the proportion of teachers receiving each rating in one school district. The performance ratings are sorted top to bottom: exemplary, proficient, needs improvement, and unsatisfactory. As demonstrated by the solid blue center of the figure, districts rate the vast majority of early career teachers as proficient. Across all teachers in this sample, 87% of teachers are identified as proficient. The limited variability in performance ratings is consistent with findings from other school systems (Kraft & Gilmour, 2016; Weisberg et al., 2009).

Figure 1 also demonstrates considerable heterogeneity in the application of performance ratings across districts. Although the state provides a model framework for evaluating teachers, the overall evaluation system retains significant elements of local decision making. State regulations permit districts to adapt the model evaluation framework or to modify existing evaluation systems to conform to its principles. Although ESE must approve district plans that depart from the model framework, Massachusetts has also structured the summative evaluation to ensure that local professional judgment plays a substantial role in determining final ratings. The model evaluation framework structures the kind of evidence used by evaluators but provides considerable flexibility in the determination of the final ratings. The summative ratings on each standard are based on a holistic review of several performance indicators rather than an aggregation of scores on individual items. Similarly, the final summative performance rating is based on the judgment of the evaluator and not on an average of performance on each of the four standards.

Figure 1. Distribution of Summative Performance Ratings by District



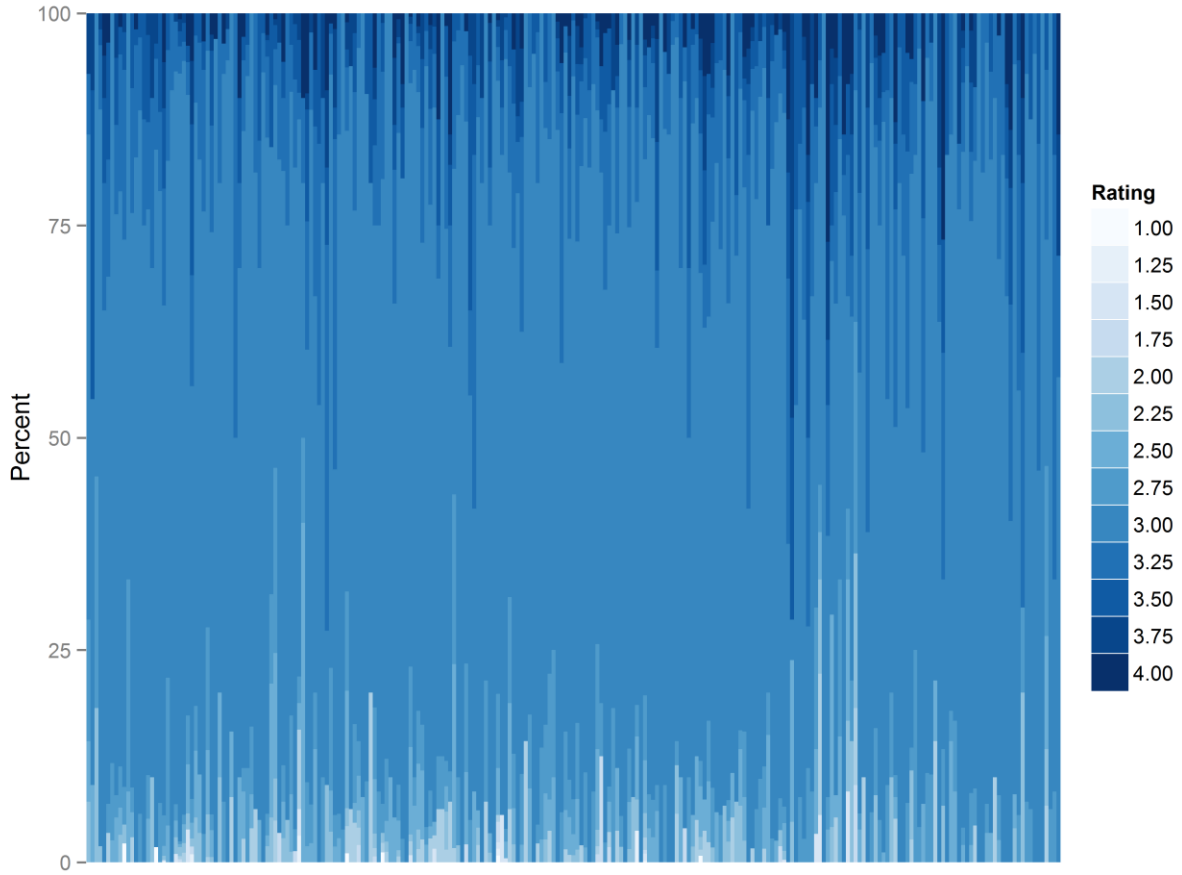
Notes: Distribution of summative performance ratings by school district. Each vertical bar indicates the distribution of ratings in a single district. We used summative performance data from our program completer sample for the 2013–2014 and 2014–2015 school years. We only plotted distributions for districts with at least 10 summative evaluations in our sample.

Consequently, we observe significant differences in the distribution of performance ratings across districts. This may pose a problem for the current study if these differences in ratings are not indicative of true differences in teacher quality due to the regional clustering of graduates near their preparation programs. Using data from the largest two districts in the sample (Boston and Springfield), we can illustrate these discrepancies in rating patterns. Among teachers with value-added data in Boston, 3% of teachers are rated as "needs improvement," 81% are rated as "proficient," and 17% are rated as "exemplary." In Springfield, 14% are rated as "needs improvement," 77% as "proficient," and 10% as "exemplary."[16] If we assume that these ratings have comparable meaning across districts, then the third percentile teacher in Boston should be

[16] These figures are similar to the overall distribution of summative performance ratings in the full sample. In Boston, 4% of teachers are rated as "needs improvement," 82% are rated as "proficient," and 13% are rated as "exemplary." In Springfield, 13% are rated as "needs improvement," 80% as "proficient," and 5% as "exemplary."

equally effective as the 14th percentile teacher in Springfield. If that were true, we would expect to see a substantial difference in average effectiveness across districts. Yet, this is not the case. Using a chi-square test, we find that teacher ratings are not equally distributed in the two districts ($p<0.01$), but we do not find the same with value added.[17] This suggests that at least some of the difference between this pair of districts is likely due to variation in the application of rating standards.[18]

Figure 2. Distribution of Aggregate Performance Ratings by District



Notes: Distribution of mean performance ratings on the four standards by school district. Each vertical bar indicates the distribution of ratings in a single district. We used summative performance data from our program completer

[17] The difference in value added between the two districts is 0.08, but not very precisely estimated. There is some evidence that teacher value added is distributed approximately normally. If we assume that Boston and Springfield both have teacher value added distributed normally with the same variance, a rough estimate of the difference in value added based on equating the 14th and 3rd percentiles of teacher effectiveness is about 0.8 teacher standard deviations. This is much less than actually observed.

[18] We also conduct a Kolmogorov-Smirnov test of differences in the distribution of value added. Both tests fail to reject the null hypothesis.

sample for the 2013–2014 and 2014–2015 school years. We only plot distributions for districts with at least 10 summative evaluations in our sample.

In light of these potential limitations in the use of the summative performance rating as a consistent statewide yardstick, we use the mean of the summative ratings on each of the four standards as our primary outcome measure.[19] The aggregated performance measure captures variation in teaching practice across different domains for a given teacher and therefore contains information beyond the final, summative measure. There is greater variability in the aggregated performance ratings constructed from the individual standard assessments. We show the distribution of these scores in **Figure 2**. The modal rating is still 3.00, which corresponds to the proficient rating and which is earned by about 70% of teachers in our sample. About 10% of teachers earn a 3.25 and another 10% earn either a 2.75 or 3.50. Second, districts and schools vary in how they map performance on the four standards into a final rating. For instance, schools may prioritize particular standards in the evaluation process and weight these more heavily in the determination of the final summative performance measure (Massachusetts Department of Elementary and Secondary Education, 2015b). The state standards require that teachers earn a proficient rating on the first two standards to earn a proficient rating overall, but districts may require different thresholds that are consistent with this requirement. We therefore average over the summative ratings on each of the four standards; however, results are similar using the reported summative rating.

Although the aggregated performance measure may provide more consistent information about teacher effectiveness across districts, the variation in district evaluation standards combined with the geographical concentration of teacher candidates near their preparation institutions complicates comparisons of institutions and programs based on their completers' summative performance ratings. Programs and institutions may be judged to be effective under such a comparison either because they graduate effective teachers or because their completers disproportionately teach in districts awarding higher ratings for a given level of performance. Because graduates of a particular program often cluster in a small number of school districts, differences in district implementation of the educator evaluation framework may represent much of the variation between programs in average ratings. We discuss our approach for accounting for differences in district evaluation standards in the Research Methods section, but, in short, most of our analyses make comparisons among teachers in a single district with different preservice preparation experiences.

Table 4. Summary Statistics for Summative Performance Samples

Program Completer Sample		Licensure Sample	
Summative Performance Rating	2.96	Summative Performance Rating	2.95
Mean Standard Score	3.03	Mean Standard Score	3.03
Enrollment	816.6	Enrollment	832.3
School % Male	51.5	School % Male	51.5
School % Hispanic	22.5	School % Hispanic	22.7

[19] The four standards contain different numbers of indicators of teaching practice and a simple mean may apply excessive weight to standards with few indicators. We conducted an exploratory factor analysis of the summative ratings. The first principal component is a nearly equally weighted average of the four standards. Consequently, analyses with the first principal component produced similar results to analyses of the mean summative rating.

Program Completer Sample		Licensure Sample	
School % Black	10.7	School % Black	10.7
School % Asian	6.0	School % Asian	5.9
School % Free Lunch	44.1	School % Free Lunch	44.0
School % Reduced Lunch	5.3	School % Reduced Lunch	5.3
School % Limited English Proficient	11.6	School % Limited English Proficient	11.4
School % Special Education	17.9	School % Special Education	18.1
Potential Experience	1.43	Potential Experience	1.45
Licensed Experience	2.09	Licensed Experience	2.07
Alternative Program	6.2	Preliminary License	40.8
Postgraduate Program	61.8	Temporary License	3.1
N	12,381	N	24,479

Note: Table presents summary statistics for summative performance rating sample. This sample includes completers in the 2010–2014 cohorts with summative performance ratings in the 2014–2015 school years. We exclude formative ratings and organizations with fewer than 15 completers. The summative rating is calculated by assigning each rating a value on a four-point scale (1=Unsatisfactory, 4=Exemplary). The mean standard score is the average summative rating on each of the four standards.

We present summary statistics for the sample of teachers with summative performance rating data in **Table 4**. We encoded the summative performance ratings on a 4-point scale with "1" indicating "unsatisfactory" and "4" indicating "exemplary." The mean summative performance rating is close to 3.0 in both the completer and licensure samples. In fact, 87% of teachers in the program completer sample and 86% of teachers in the licensure sample earned the proficient rating. An additional 5% of teachers earn the exemplary rating in each sample. The average aggregate rating, which takes the mean of the ratings on the four standards, is also similar for both samples. Looking at teachers' preservice experiences, we see that the distribution of teacher entry pathways is similar to the student achievement samples. Nearly 62% of teachers completed a postgraduate program, and 6% of teachers completed an alternative program. In the licensure sample, about 57% of teachers first qualified for an initial license, 40% for a preliminary license, and 3% for a temporary license.

Teacher Attrition Data

For the analysis of teacher attrition, we use a sample of teachers working in full-time teaching roles in Massachusetts. We made two additional restrictions to the sample for this analysis. We exclude from the sample teachers who remain in the workforce but move to a new role, such as administrator, or are not working as full-time employees. In the terminology of duration models, these observations are *censored*; these teachers end their time as classroom teachers without exiting the profession. Because we do not observe what would have happened had they remained in the classroom, we omit these years from the analysis. Such teachers therefore contribute to our estimates of program attrition rates only while acting as classroom teachers. We additionally exclude all observations in 2015 from this analysis because we do not observe whether teachers return to Massachusetts public schools in 2016. The exclusion of the 2015 data also necessitates the exclusion of the 2014 cohort from our analysis.

Table 5. Summary Statistics for Attrition Samples

Program Completer Sample		Licensure Sample	
Percent Exiting	7.0	Percent Exiting	9.3
Enrollment	864.0	Enrollment	861.5
School % Male	51.4	School % Male	51.4
School % Hispanic	20.3	School % Hispanic	20.9
School % Black	11.0	School % Black	11.1
School % Asian	5.7	School % Asian	5.5
School % Free Lunch	40.2	School % Free Lunch	40.1
School % Reduced Lunch	6.2	School % Reduced Lunch	6.2
School % Limited English Proficient	9.9	School % Limited English Proficient	10.0
School % Special Education	17.9	School % Special Education	18.1
Observed Experience	0.94	Observed Experience	0.99
Pct. Alternative Program	7.0	Pct. Preliminary License	40.8
Pct. Post-BA Program	67.0	Pct. Temporary License	3.1
N	17,661	N	33,564

Note: Table presents summary statistics for attrition sample. This sample includes completers in the 2010–2013 cohorts working in teaching roles during the 2011–2014 school years. We excluded former teachers in non-teaching roles, teachers who have previously left Massachusetts public schools, and organizations with fewer than 15 completers. Abbreviations: Exp = experience; Prog = program, BA = bachelor's. Standard deviations in parentheses.

These restrictions leave us with a sample of 7,640 teachers (and 17,661 teacher-year observations) linked to program completions and 13,991 teachers (and 33,564 teacher-year observations) linked to first-time teaching licenses. We describe this sample in **Table 5**. The sample included teachers from the 2010–2013 cohorts during the 2011–2014 school years and tracks teachers for up to 4 years following their first completion or license. Among program completers in Massachusetts, about 7% of teachers leave the Massachusetts public school system each year. In the licensure sample, which includes teachers with out-of-state credentials and those who have not completed an educator preparation program, the attrition rate is about 9% each year. In the attrition analysis, we used the observed experience measure in order to enumerate the years in the current teaching spell. In both samples, the average teacher had slightly less than 1 year of teaching experience.[20]

Research Methods

As described in the introduction, our analysis focuses on three primary research questions:

[20] Recall that the data are organized at the teacher-year level. Because we observe more recent cohorts for fewer total years, there are disproportionately many novice teachers in the sample. The average teacher experience for observations in the sample is therefore less than the average experience that will be attained by teachers in the sample.

1. What is the variation in student achievement gains associated with different teacher licensure pathways, preparation program types, and specific educator preparation institutions and programs?

2. What is the variation in teacher evaluation results associated with different teacher licensure pathways, preparation program types, and specific educator preparation institutions and programs?

3. What is the variation in teacher retention associated with different teacher licensure pathways, preparation program types, and specific educator preparation institutions and programs?

These three questions are intended to provide a preliminary description of program completers' performance using a variety of important measures that have been linked directly to student learning. For each of these research questions, we focus our analysis on three different descriptors of teachers' preservice preparation. The first descriptor is a teacher's route into the profession, which we classified on the basis of the type of license a teacher possesses. We then consider differences in outcomes by program type among those completing an educator preparation program in Massachusetts.

We next focus on differences in educator effectiveness across educator preparation institutions within each licensure pathway. The purpose of this analysis is twofold. First, we estimate individual contrasts in educator effectiveness and retention rates for particular institutions. Such contrasts help us understand the variation between educator preparation programs. Second, we quantify the variability in educator outcomes across institutions by estimating the variance in teacher outcomes across institutions. This latter measure provides a summary of the variability in outcomes for graduates of different institutions across the state and provides some context for whether the differences in outcomes we observed are either substantively or statistically important.

Finally, educator preparation institutions often offer programs for both undergraduate and graduate students and typically train teachers for a number of different teaching roles through separate programs, and these different programs may have different curricula, admissions, or student teaching requirements and may therefore produce teachers with differing average outcomes. We therefore conclude by focusing within institutions and estimating the variation in teacher outcomes across different programs within the same institution. Given that disaggregating results by program produces many fewer teacher observations per preparation program, our objective is only to quantify the magnitude of the variability in outcomes across programs and not to estimate individual contrasts in educator effectiveness and retention rates for particular programs.

We provide methodological details for each component of these analyses in the following subsections. Because of the similarities in the analyses addressing different research questions using the same data, we group the methodological discussion by data set.

Educator Preparation and Student Achievement

Our analysis of the effects of teachers with various educator preparation backgrounds on student achievement uses data from student performance on the MCAS and PARCC assessments. In order to separate the influence of teachers with varying backgrounds from other student and school factors, we estimate a set of value-added models. Value-added models assess teachers' contributions to student learning by focusing on changes in achievement test scores from one year to the next. By examining changes in student achievement instead of student achievement levels, we may better distinguish the effects of teachers from family background and other schooling factors that influence learning.

Researchers have used value-added methods to compare the effectiveness of teachers with varying characteristics in several states and contexts. For example, researchers have used methods similar to those used in this study to estimate how much more productive teachers become as they gain experience (Papay & Kraft, 2015; Wiswall, 2013), the relative effectiveness of teachers without standard teaching licenses (Boyd et al., 2006; Clotfelter et al., 2006; Goldhaber & Brewer, 2000; Kane et al., 2008), the effectiveness of National Board-certified teachers (Cantrell et al., 2008; Clotfelter et al., 2006; Cowan & Goldhaber, 2016; Harris & Sass, 2009), and the degree to which teachers specialize in particular subjects or grade levels (Cook & Mansfield, 2016; Condie et al., 2014; Goldhaber et al., 2013a; Ost, 2014). In a series of papers spanning several states, researchers have also used these methods to compare the effectiveness of teachers from different educator preparation programs (Boyd et al., 2009; Goldhaber et al., 2013b; Henry et al., 2014; Koedel et al., 2015b; Mihaly et al., 2013; von Hippel et al., 2016).

We standardize all student test scores prior to estimation so that the estimates of program and pathway effects represent differences in student learning in terms of standard deviations in the student achievement distribution. To provide some context for the meaning of estimates expressed in these terms, it is useful to establish some benchmarks for the magnitude of the influence of teachers on student outcomes. Prior research has estimated that one standard deviation in the distribution of teacher effectiveness—or the difference between a 50^{th} percentile teacher and a teacher at the 67^{th} percentile of the distribution— corresponds to about 0.10 to 0.25 standard deviations in the student achievement distribution (Aaronson et al., 2006; Chetty et al., 2014a; Goldhaber et al., 2013c; Jackson, 2013). According to research by Chetty et al. (2014b), exposure to a teacher with one standard deviation higher value added raises average annual earnings at age 28 by about $350. Teacher credentials also provide possible benchmarks for achievement effects. Alternatively, we could describe variation in teacher quality in terms of average annual student learning. Using nationally normed assessments, Bloom et al. (2008) estimate that average annual learning gains in the grades we consider amount to 0.29 student standard deviations in reading and 0.39 standard deviations in math.

We estimate value-added models using a specification similar to those estimated in this broader literature. Our most basic models include controls for prior year test scores in math and ELA, student demographic characteristics, indicators for participation in special programs, and the

classroom and school means of these characteristics.[21] These models assume that classroom assignments do not vary systematically with unobserved factors that affect student achievement.[22] At the elementary and middle school level, recent analyses by Chetty et al. (2014a) and Bacher-Hicks et al. (2014) found that models similar to ours predicted teacher effectiveness with little or no bias. However, this remains a controversial assumption and some analyses suggest that value-added models may contain a substantial amount of bias (Rothstein, 2010). In order to test the sensitivity of our results to the possibility of student sorting, we estimate several additional models.

Our first robustness check investigates the sensitivity of our results to more robust controls for school context. Prior research has found that teachers are more likely to obtain positions in schools near their hometowns and preparation programs (Boyd et al., 2005; Krieg et al., 2016). The geographic clustering of program completers may bias our estimates if certain schools or regions differ in their effectiveness. We therefore estimate models that include school fixed effects. School fixed effects models base the estimation of program and pathway effects on within-school variation in student achievement. This approach compares teachers from one program to other teachers from different programs who are working in the same school. These models account for the possibility that programs send their graduates to particular schools that may differ in school leadership, financial resources, or other unobserved ways.

This approach comes with two limitations, both discussed more fully by Mihaly et al. (2013). First, by restricting the sample of comparison teachers to those in the same school, fixed effects methods are considerably less precisely estimated than the baseline models.[23] Second, school fixed effects models may overcontrol for real differences in teacher effectiveness across schools. These models adjust for both unobserved school level factors (such as the quality of leadership, the amount of funding per pupil, the contributions of non-teaching staff) and the average level of teacher effectiveness in the school. Therefore, programs that tend to place their graduates in schools with graduates of more effective programs will have lower estimates in school fixed effects models. If schools tend to hire teachers of similar quality, so that graduates of more effective programs are more likely to work in schools with completers of other more effective programs, this will result in a more compressed distribution of estimated program effectiveness.

[21] In particular, we include a cubic polynomial in prior year test scores, student race/ethnicity, subsidized lunch status, limited English proficiency, special education status, and school and classroom means of these variables. For 10th grade students, we use eighth grade test scores. The empirical methodology is explained in greater detail in Appendix A.

[22] These models are the most similar conceptually to the student growth percentiles (SGP) models estimated as part of the educator effectiveness framework in Massachusetts. Prior research has found that these two methods produce similar estimates of teacher effectiveness (Goldhaber et al., 2013c), but there are some differences between the two models in the specification of the relationship between present and past achievement and in which student and classroom controls are included. In Appendix A, we show where results from the two models differ.

[23] The precision of the estimates will depend on how frequently we observe teachers from multiple programs in the same school. In our value-added samples, each of the programs is connected in the sense described by Mihaly et al. (2013) so that program effects are identified relative to a common mean in the school fixed effects models. However, given the short period available for estimation, many graduates are not observed in schools with other recent completers. In both the math and ELA samples, 49% of schools hire completers from only a single institution into the testing data during the time period we consider.

Despite these limitations, school fixed effects models may be more robust when there are differences in the student populations or types of schools served by institutions. Estimates that significantly diverge from the baseline models may indicate programs for which traditional approaches yield poor estimates of true effectiveness. In addition, some of the institutions we studied are connected to charter schools or particular school districts. For instance, some of the charter authorizers administer teacher education programs that predominantly feed into their own schools. Studies of some of these specific charter school networks in Massachusetts suggest they are more effective at producing student learning gains (Angrist et al., 2016). If these schools are better at raising student achievement for reasons other than their teaching staff, then those differences in schooling context will be reflected in the estimates of the teacher preparation institution's average value added. In these cases, it is difficult to disentangle the effects of the school environment from the effectiveness of the teaching staff.

The basic school fixed effects model accounts for cross-school sorting of students by making within-school comparisons of teachers with varying backgrounds. They do not control for within-school sorting or tracking. Classroom assignment policies may bias our estimates if principals group students based on unobservable characteristics and graduates from particular programs are more or less likely to teach in favorable classrooms. For instance, some programs may specialize in more advanced math or science fields that are taken disproportionately by academically advanced students. These students may also perform well on state tests because they have access to other resources either at school or home. For the licensure and program pathway analyses, we estimate models that replaced controls for prior test scores with student fixed effects. The intuition behind these models is similar to that for the school fixed effects methods. We compare student achievement for years in which students have a teacher from a particular pathway to other years in which the same student has a teacher from another pathway. Using this within-student variation in teacher staffing removes time-invariant, unobserved student characteristics that may be associated with assignment to teachers from different licensure pathways or program types.[24]

Educator Preparation and Performance Evaluations

The summative performance ratings are aligned with Massachusetts' expectations for professional teaching practice and therefore provide a useful metric to measure teacher effectiveness. As we discuss in the data section, the summative ratings are available for a larger set of teachers and may provide information about effective teaching practice that is not captured by student standardized test scores. The challenge is extracting the signal about preparation experiences and teacher effectiveness from variation in the implementation of the evaluation framework. The simplest approach would be to calculate mean summative ratings for each institution and pathway. However, the unadjusted summative ratings may not provide a common measure of teacher effectiveness in all school districts and this approach may not produce unbiased estimates of differences in effectiveness across institutions and pathways. We therefore

[24] We do not pursue this approach with the program and institution indicators because of the substantial increase in the estimation error variance. Instead, we estimate models that control for the tracks within schools based on the academic level of the course (e.g., remedial, general, advanced) using methods suggested by Jackson (2014) and Protik et al. (2013). Because the academic level information may not be consistently defined across school systems, we generate track effects for each school. We include results from these models in Appendix A.

use the performance measures to estimate several models designed to adjust for these confounding factors.

Our first adjustment accounts for the influence of school context and teaching assignments on teachers' ratings using observable characteristics of teachers' schools. Our use of this method is guided by research on classroom observations of teaching practice, which are one component of the final assessment measure. Researchers have found that these observational measures are correlated with student characteristics, which may indicate that some teachers receive more favorable evaluations because of the students they teach (Steinberg & Garrett, 2016; Whitehurst et al., 2014). We therefore estimate models that controlled for the aggregated student characteristics included in the value-added models. As with the value-added models, we also include controls for teacher experience and school year in order to adjust for differences in the proportion of teachers graduating in particular years.[25] These models rely on similar assumptions as our basic value-added models. In particular, we assume that unobserved factors affecting teacher ratings, such as district implementation, classroom dynamics, or school leadership, are not correlated with a teacher's preparation program.

In Figures 2 and 3, we present graphical evidence that contradicts this assumption. Specifically, the rating scales appear to differ by school district. The discrepancies in the distribution of ratings may reflect the role of local decision makers in assigning ratings under the educator evaluation framework. Given the regional segmentation of teacher labor markets, these idiosyncrasies in local scoring are likely to be reflected in estimates of program effects. We therefore follow an approach suggested by Ronfeldt and Campbell (2016) and estimate models that included school or district fixed effects. The fixed effects models serve a similar purpose as in the value-added models, where we use them to control for unobserved characteristics of a school that are shared by all students and teachers. In this case, we compare teachers from one organization to teachers from other organizations in the same district (or school). By making comparisons within district, we remove variation in mean ratings across districts that may be caused by differences in the implementation of the evaluation framework. As is the case with the value added models, this comes at the expense of precision in the estimates.

Educator Preparation and Retention

The final outcome we consider is teacher retention in Massachusetts public schools. We follow a well-established literature in the economic analysis of teacher labor markets and estimate a discrete time hazard model that follows teachers during their first teaching spell in Massachusetts public schools following licensure or program completion. We estimate the probability that a teacher leaves the Massachusetts public school system at the conclusion of each school year conditional on preservice preparation experiences. Given the cohorts we consider and the data available, the analysis follows teachers for up to 4 years after licensure or program completion. Similar models have been used to study the career pathways of teachers with different

[25] We discuss the estimation procedures more fully in Appendix A. For the models presented in the text, we controlled for the potential experience measure that count years since program completion. All of the models estimated in the text are implemented as linear regression models. Although the data are discrete and we uncovered some evidence that the sensitivity of the rating scale to teacher quality differs across districts, we show in the appendix that the choice of link function is much less important empirically than assumptions about the correlation of educator preparation institution indicators with school or district unobservables.

preparation backgrounds and the influence of a number of school and district characteristics on teacher retention (Clotfelter et al., 2008; Goldhaber & Cowan, 2014; Goldhaber et al., 2011, 2016; Imazeki, 2005; Kane et al. 2008). For each of the attrition analyses, we consider teachers during their first spell in Massachusetts public schools. We therefore drop teachers once they transitioned into a part-time or non-teaching role. These teachers do contribute to our estimation for the years in which they teach, but attrition that follows these transitions is not considered. Similarly, we do not consider teachers' subsequent spells following a break in service. Although this is consistent with prior research, nearly 25% of new teachers each year are re-entering the profession (Provasnik & Dorfman, 2005).[26]

Our most basic models are similar to the baseline models for the analysis of summative performance ratings. We estimate regression models where the dependent variable is an indicator for whether a teacher leaves the public school system and include controls for experience, school characteristics, and school year.[27] The assumptions necessary for the estimates from this model to provide unbiased institution indicators are therefore similar to those we discussed previously. In particular, it must be the case that unobserved factors associated with staff turnover are not associated with teachers' preparation programs. As with the summative assessment data, there are several reasons to believe this is not strictly true. Previous analyses have demonstrated several school-level factors that are associated with staff turnover. These include salary (Clotfelter et al., 2008; Springer et al., 2016), school leadership (Branch et al., 2012; Boyd et al., 2011; Jacob, 2011), mentoring (Rockoff, 2008; Smith & Ingersoll, 2004), and the availability of other nearby teaching positions (Imazeki, 2005; Jackson, 2012). It is less clear that these or other factors vary systematically with teachers' preparation background, although we do observe differences across sponsoring organizations in school characteristics in Appendix A, which should provide some reason for concern. Because this may not be a reasonable assumption, we also follow the approach of Goldhaber and Cowan (2014), who relax the assumption that school-level unobservables are unrelated to the composition of the teaching staff.[28] These models base comparisons of teacher retention on within-school comparisons to teachers from other backgrounds and are similar to the models we estimate for the student achievement and summative performance data.

Results: Teacher Preparation and Effectiveness

In this study, we assess variation in teacher effectiveness across licensure pathways and educator preparation programs using two primary measures of teacher effectiveness (aligned with research

[26] Given that the Massachusetts data do not include information on employment in other school systems or assignments before 2010, we cannot compute exactly comparable figures. However, over the period 2013-2015, 25.6% of new employees working as teachers in Massachusetts public schools have some prior experience after 2010.

[27] We discuss the model more fully in Appendix A. We implement all of the attrition models as probit models. Because we are estimating duration models, we use the observed experience measure, which counts years in the public school system following licensure or completion.

[28] We provide additional details in the appendix. We estimated a school correlated random effects model following a parameterization similar to that suggested by Wooldridge (2010). We included school-level means of each of the observation-varying variables in the model (including the staffing characteristics) so that identification of organization and pathway effects is based on comparisons of teachers with different programs/pathways who work in schools with the same mean staffing profile. Goldhaber and Cowan (2014) showed that this approach produces estimates very similar to linear models with school or school-by-year fixed effects.

questions 1 and 2). We begin by describing how test-based measures reflect educators' backgrounds using value-added measures. The value-added measures are based on student performance on the MCAS and PARCC assessments. We then proceed to an analysis of the variation in teacher performance on the Massachusetts educator evaluation framework. In our analysis of teacher evaluation data, we focus on the summative performance rating, which is the culmination of the teacher evaluation process.

Each of these measures has advantages and disadvantages. An advantage of the value-added analysis is that the state standardized tests measure student achievement in a uniform way so that differences in performance measures should have a common meaning across districts. This is not necessarily the case with the summative performance data given that the evaluation framework provides wide latitude for local decision making and the professional judgment of individual evaluators. The value-added measures also provide a finer measure of teacher performance. As we have seen, most teachers receive the proficient rating on the summative measure. On the other hand, subjective evaluations contain information about teacher effectiveness that is not well captured by value-added assessment. Test-based value-added measures may not be highly correlated with teachers' effects on student non-test outcomes or with their contributions to school leadership or climate (Gershenson, 2016; Harris & Sass, 2014; Jackson, 2016). Evaluators may be able to provide a better assessment of these teaching skills. Moreover, value-added ratings are only available for teachers in tested grades and subjects, whereas the summative performance ratings are available for a much larger group of teachers.

Preparation Pathways and Student Achievement (RQ 1)

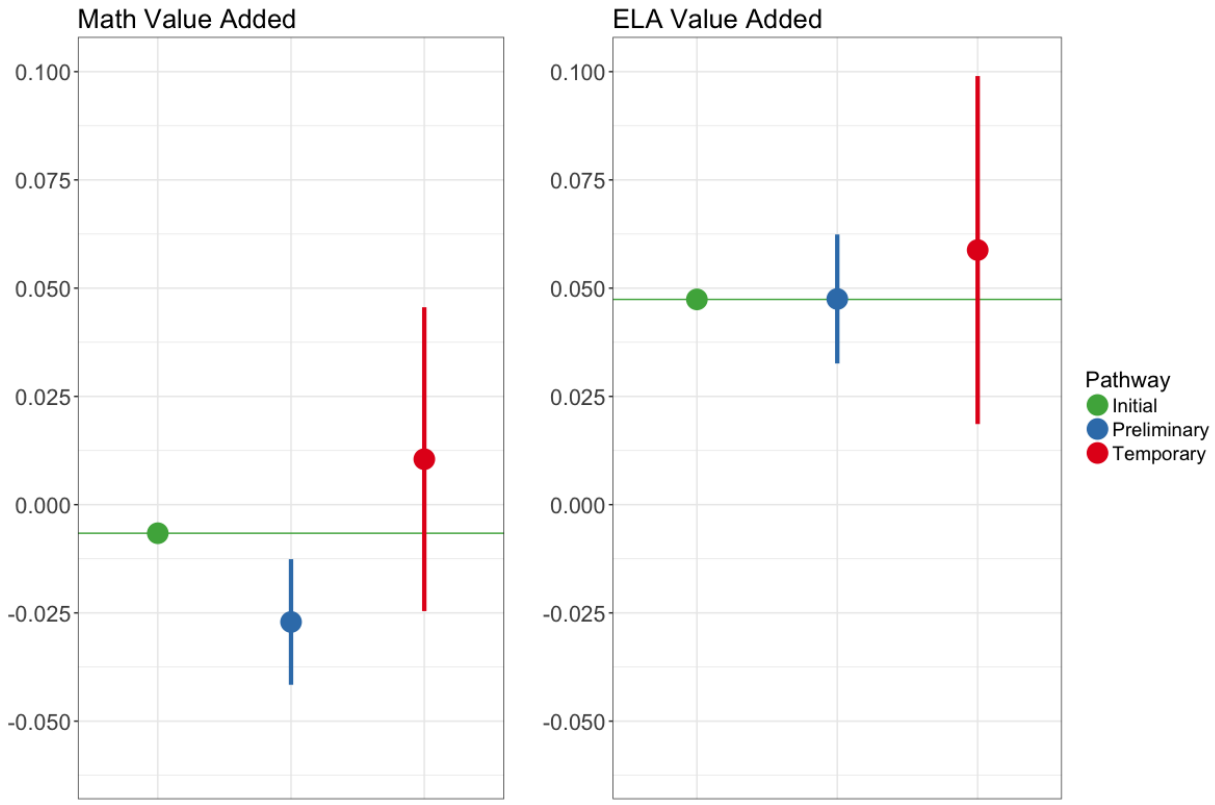
We assess variation in student achievement (research question #1) by several descriptors of teacher preservice preparation. The most general is by the type of license a teacher first earns in Massachusetts. Most teachers enter the state's public teaching workforce after having completed an educator preparation program within the state, but there are also options for teachers arriving from outside the state or for prospective teachers who have yet to complete an educator preparation program. We then examine teachers who have completed a Massachusetts educator preparation program and compare teacher effectiveness across undergraduate, graduate, and alternative programs. Looking across the value-added measures, we generally find mixed evidence of licensure or program type effects. Finally, we estimate the average effectiveness of graduates from individual institutions in Massachusetts. Although we find consistent evidence of differences among individual institutions, given the small number of teachers graduating from many programs, we can only distinguish a few individual institutions from the state average. Finally, across all levels of aggregation, we find evidence for a general truism in the educator effectiveness literature: there is far more variation in teacher quality within categories of teachers than across them.

Licensure Pathways and Student Achievement

We begin with an analysis of teachers' value added by their first teaching license type in Massachusetts. We consider three license types: initial, which is offered to teachers who have completed a preparation program and state testing requirements; preliminary, which is offered to teachers who have not completed an approved program but have completed state testing

requirements; and temporary, which is offered to teachers with at least 3 years of experience in another state but who have not completed state licensure testing requirements.

Figure 3. Student Achievement by Licensure Pathway



Note: Preliminary (temporary) license indicates that a teacher's first Massachusetts license is of the preliminary (temporary) type. Teachers who enter with an initial license comprise the omitted group. Estimated effects are derived from the value-added models described in the text. All models include controls for teacher experience, student, class, and school demographics, and year and grade effects. Predicted values are estimated at the mean of the student characteristics. Standard account for clustering at the teacher level. *N* = 282,083 (math); *N* = 246,971 (ELA). ELA = English Language Arts.

Teachers who enter with the initial and temporary license types have all completed educator preparation programs before they begin teaching. These groups are therefore likely to be most similar in terms of prior preparation for student teaching. Perhaps unsurprisingly, therefore, we find little evidence of differences in teaching effectiveness across these two licensure pathways. The basic results are summarized in **Figure 3**. After estimating our value-added models, we plot the predicted student achievement by licensure pathway. The points indicate predictions for student achievement and the solid lines indicate 95% confidence intervals. Points that exclude the horizontal green line are statistically significantly different than the estimates for teachers with initial licenses. In math, we find that teachers with preliminary licenses produce lower achievement gains than those with initial licenses or temporary licenses. On the other hand, in ELA, there is little evidence that teacher effectiveness varies across licensure pathway. The estimated achievement levels are all quite similar and none is statistically distinguishable from the others.

We present the results more formally in **Table 6**. The estimated coefficients on preliminary and temporary license types now compare the average effectiveness of each of these groups to teachers who first earn an initial license (the reference category). In the first row of Table 6, we compare teachers with preliminary licenses to those entering with initial licenses. Here, results differ across subjects. In math, we find that teachers with preliminary licenses are about 0.02 standard deviations less effective than teachers with initial licenses. The result is statistically significant in our baseline model as well as models with school fixed effects. It is not significant in the student fixed effects model, although the coefficient is nearly identical to the other results. In ELA, on the other hand, we find little evidence that teachers with preliminary licenses are differentially effective than teachers with initial licenses.

Table 6. Average Differences in Student Achievement Relative to Teachers with Initial Licenses

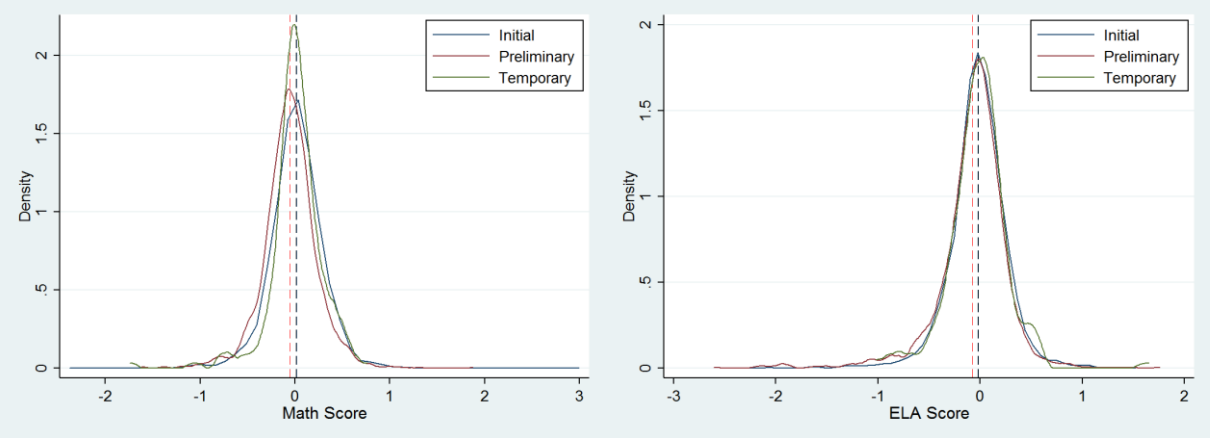
	Math			ELA		
Licensure Pathway	**Baseline Model (1)**	**School Fixed Effects (2)**	**Student Fixed Effects (3)**	**Baseline Model (4)**	**School Fixed Effects (5)**	**Student Fixed Effects (6)**
Preliminary License	-0.020*** (0.007)	-0.018*** (0.007)	-0.021 (0.015)	0.000 (0.008)	0.009 (0.007)	-0.008 (0.018)
Temporary License	0.017 (0.018)	0.003 (0.013)	0.017 (0.033)	0.011 (0.020)	-0.026* (0.015)	-0.028 (0.059)

Note: Preliminary (temporary) license indicates that a teacher's first Massachusetts license is of the preliminary (temporary) type. Teachers who enter with an initial license comprise the omitted group. Estimated effects are derived from the value-added models described in the text. All models include controls for teacher experience, student, class, and school demographics, and year and grade effects. The models in columns 2 and 3 control for school and student fixed effects, respectively. Standard errors (in parentheses) account for clustering at the teacher level. N = 282,083 (math); N = 246,971 (ELA). *p < 0.10, **p < 0.05, ***p < 0.01. ELA = English Language Arts.

In the second row, we compare the effectiveness of teachers entering with initial and temporary licenses. Looking across math and ELA models, the coefficients on the temporary license group are neither consistently positive nor negative and there are no statistically significant results at the 5% level. These results suggest there is not a significant difference between teachers with initial and temporary licenses. This result is somewhat at odds with an analysis by Bastian & Henry (2015), who found that out-of-state teachers were somewhat less effective than teachers trained in North Carolina, but the finding is generally consistent with the findings of Goldhaber et al. (2013b), who found that most programs in Washington state produced graduates who are not statistically distinguishable from out-of-state teachers. One potential point of contrast with these prior studies is that we do not observe teacher experience directly in the Massachusetts data. Databases in Washington and North Carolina provide measures of experience derived from payroll records that note the location of a teacher on salary schedules. In these models, we measured experience as the time elapsed from the license issue date. Given the minimum experience requirements for teachers entering with temporary licenses, it is likely that teachers with temporary licenses have more unobserved teaching experience than teachers with initial licenses. It may therefore be the case that true experience-adjusted models would find them to be less effective teachers. Regardless of this possibility, it remains the case that first-year teachers in Massachusetts with temporary licenses do not appear differentially effective than those with initial licenses.

The ambiguity of the comparison between teachers with preliminary licenses and initial licenses is consistent with the broader literature that compares traditional entry pathways to alternative licenses that do not require the completion of a traditional educator preparation program. In **Figure 4**, we demonstrate one commonality with the previous literature: most of the variation in teacher quality exists within, rather than between, licensure pathways. We plot the distribution of teacher value added for each of the three pathways on the same plot separately for math and ELA. The vertical axis denotes the density of value added for the value given on the horizontal axis and is an indication of the likelihood of observing a teacher with a given value added estimate. Higher values indicate a greater likelihood that teachers with nearby value-added scores are observed in the data. Groups with more effective teachers have curves that are shifted to the right. Therefore, more similar populations will have greater overlap in their density curves and less similar populations will be more spread out along the horizontal axis. Despite the small mean differences in the curves plotted for each pathway, there is a great deal of overlap in the distributions. In other words, the fact that the density curves lie close to each other indicates that the probability of observing a teacher of a given effectiveness level is fairly similar in each of the licensure pathways. Even in cases where the licensure pathway provides information on *average* effectiveness, it is likely to provide a poor prediction for the effectiveness of an *individual* teacher.

Figure 4. Distribution of Value Added by License Type



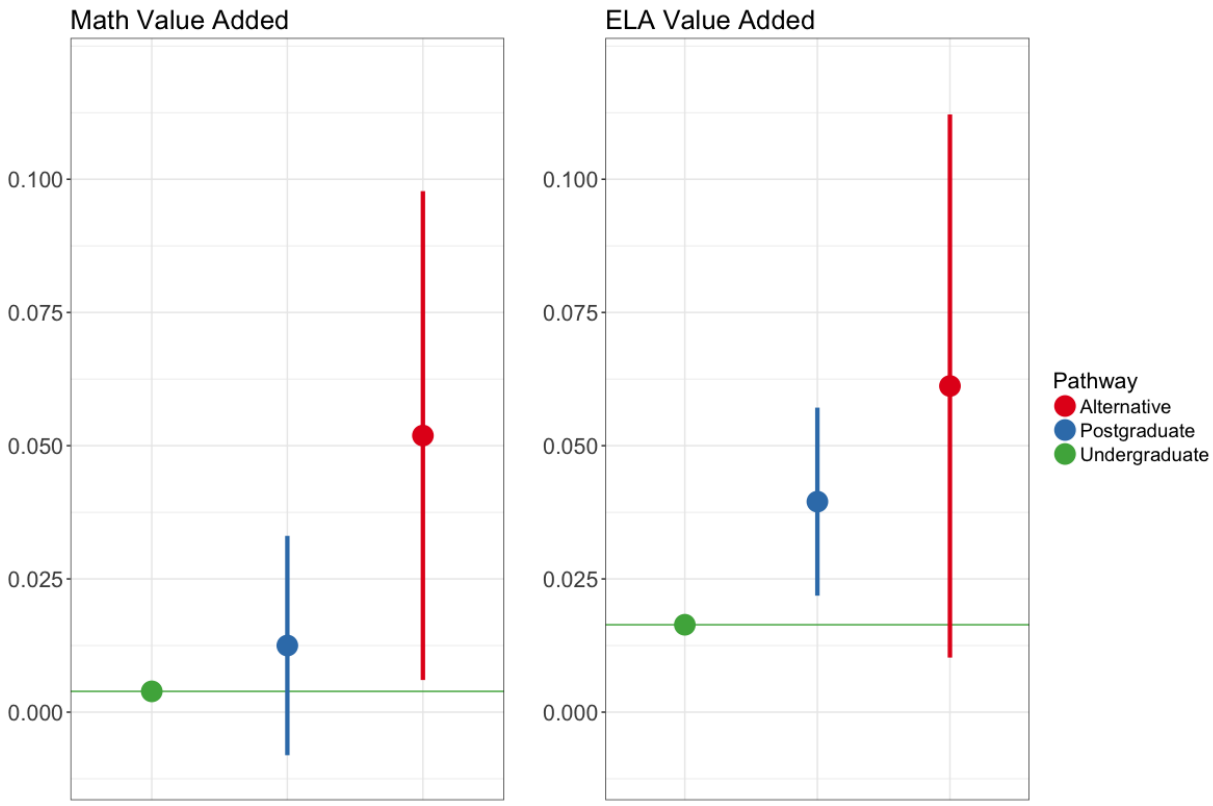
Note: Figure displays kernel density estimates of teacher value-added distribution in math and ELA by first license type. Dashed lines indicate license type means. The value-added models include the same control variables included in the regression analyses with the exception of license type indicators. We estimated a single teacher effect across all years of available data for each teacher in the sample. ELA = English Language Arts.

Program Type and Student Achievement

In the remaining analyses, we limit our sample to teachers we could successfully link to an educator preparation program in Massachusetts and an associated licensure record following completion. Before moving on to the individual institution and program contrasts, we examine differences in the average effectiveness of program completers by program type. We again summarize the basic results by showing the predicted student achievement for students assigned teachers from various program pathways. These are displayed in **Figure 5**. Students who are assigned teachers from alternative programs have higher achievement than those assigned

teachers from other pathways. In the baseline models shown in Figure 5, this is significant in math but not ELA; however, we demonstrate below that this finding is somewhat sensitive to modeling decisions. We also find that teachers from postgraduate programs are more effective than those from undergraduate programs in ELA. In this case, the magnitude of this result is robust to the various specification checks described below.

Figure 5. Student Achievement by Program Pathway



Note: Postgraduate (alternative) program indicates the type of Massachusetts program a teacher attended. Teachers who attended an undergraduate program comprise the omitted group. Estimated effects are derived from the value-added models described in the text. All models include controls for teacher experience, student, class, and school demographics, and year and grade effects. Predicted values are estimated at the mean of the student characteristics. Standard errors account for clustering at the teacher level. *N* = 141,664 (math); *N* = 128,219 (ELA). ELA = English Language Arts.

The results in **Table 7** use similar models as in the licensure pathway results above. In this case, teachers from undergraduate institutions comprise the reference category. Therefore, the coefficients on postgraduate and alternative programs represent contrasts with teachers from undergraduate programs. We find some evidence that graduates of postgraduate programs are more effective than graduates of baccalaureate programs, but the results are not consistent across subjects. In math, we find no evidence of such an effect. The point estimates are small and statistically insignificant. On the other hand, we find that teachers from postgraduate programs are more effective in ELA classrooms. The point estimates range between 0.02 and 0.03 student standard deviations. The estimates are statistically significant in both the baseline and school fixed effects models; although not significant in the student fixed effects model, the point

estimates are quite similar. These estimated contrasts are similar in magnitude to the difference between teachers with National Board certification and those without (Clotfelter et al., 2007; Cowan & Goldhaber, 2016; Goldhaber & Anthony, 2007). They are also more robust across models than is typical of assessments of the value of graduate degrees: previous analyses of student achievement data have typically not found that teachers with masters' degrees are more effective than those without. However, these analyses are generally based on teachers' current level of educational attainment, and not on the type of education preparation program a teacher attended. In a more comparable analysis, Henry et al. (2014) found that high school math teachers in North Carolina who completed an in-state graduate program were more effective than undergraduates; however, this result did not hold for other levels.

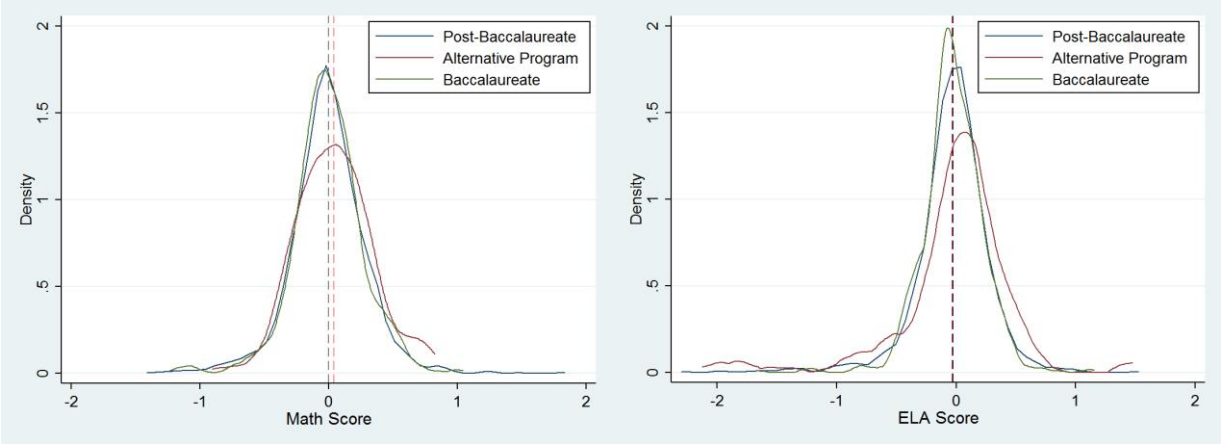
Table 7. Average Differences in Student Achievement Relative to Teachers from Undergraduate Programs

	Math			ELA		
Program Type	**Baseline Model (1)**	**School Fixed Effects (2)**	**Student Fixed Effects (3)**	**Baseline Model (4)**	**School Fixed Effects (5)**	**Student Fixed Effects (6)**
Postgraduate Program	0.009 (0.010)	0.007 (0.010)	-0.019 (0.037)	0.023** (0.010)	0.023** (0.010)	0.033 (0.038)
Alternative Program	0.048** (0.023)	0.045* (0.023)	0.008 (0.071)	0.045* (0.026)	-0.016 (0.031)	0.041 (0.088)

Note: Postgraduate (alternative) program indicates the type of Massachusetts program a teacher attended. Teachers who attended an undergraduate program comprise the omitted group. Estimated effects are derived from the value-added models described in the text. All models include controls for teacher experience, student, class, and school demographics, and year and grade effects. The models in columns 2 and 3 control for school and student fixed effects, respectively. Standard errors (in parentheses) account for clustering at the teacher level. N = 141,664 (math); N = 128,219 (ELA). *p < 0.10, **p < 0.05, ***p < 0.01. ELA = English Language Arts.

There is less evidence of differences between undergraduate programs and alternative programs. In math, our estimates are generally positive, albeit only statistically significant in the baseline model. The estimated effects are also sensitive to the inclusion of school and student fixed effects. In ELA, none of the estimates is statistically significant and the sign of the effects is also mixed. Interestingly, the pattern of results in both math and ELA, with significantly higher results in the baseline model than in the fixed effects models, suggests that teachers from alternative programs work in environments with favorable unobserved determinants of student achievement. One possible explanation is the connection between alternative programs and some high-performing charter networks (Angrist et al., 2016). If completers of alternative programs tend to teach in high-performing schools, then we would expect to see smaller differences when making the within-school and within-student comparisons in columns (2) to (3) and (5) to (6).

Figure 6. Distribution of Value Added by Program Type



Note: Figure displays kernel density estimates of teacher value-added distribution in math and ELA by program type. Dashed lines indicate program type means. The value-added models include the same control variables included in the regression analyses with the exception of program type indicators. We estimated a single teacher effect across all years of available data for each teacher in the sample. ELA = English Language Arts.

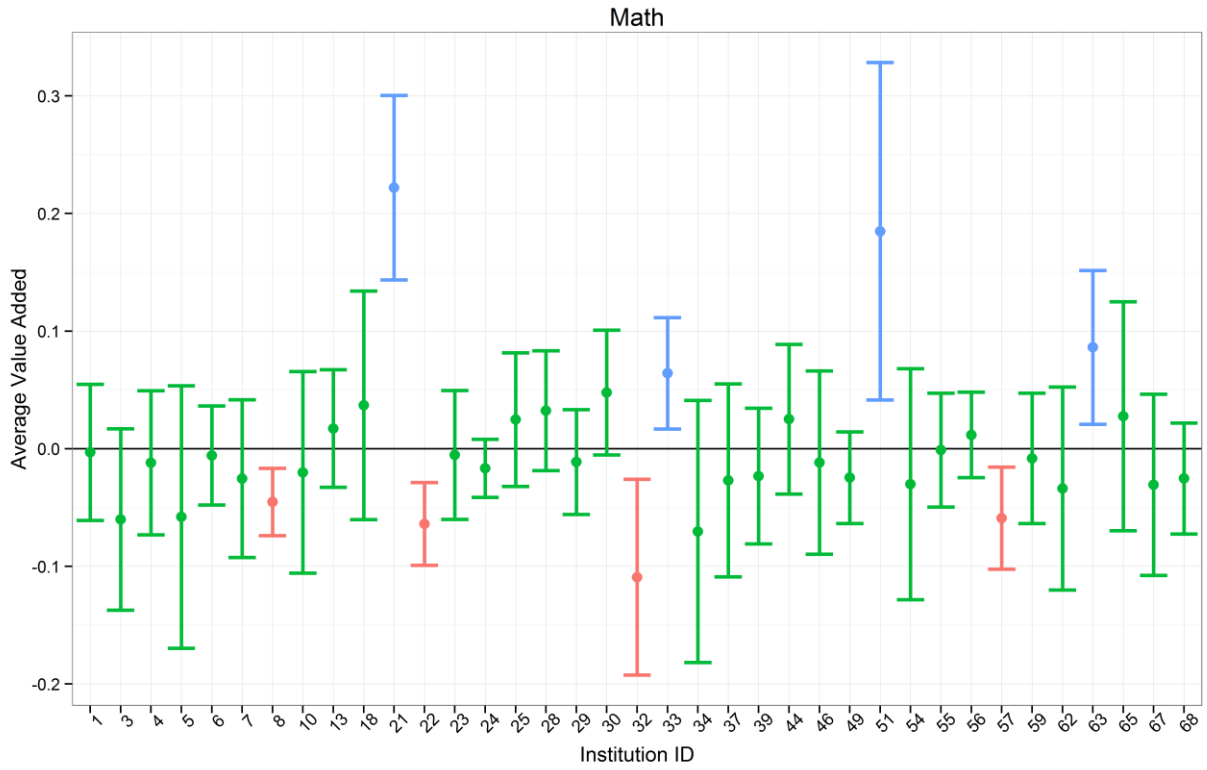
As with the licensure pathway analysis, we can depict the relative variation in teacher and pathway effects graphically by plotting the value-added distribution separately by program type. The density plots in **Figure 6** indicate that there is substantial overlap in the distributions of teacher quality by program type. The differences in mean effectiveness evident in the baseline models from Table 7 are apparent in the figures; however, the overall distribution of teacher quality is fairly similar across program types. As most analyses have found, there is a mix of both effective and ineffective teachers in each program type.

Programs, Institutions, and Student Achievement

We next replace the pathway variables with indicators for each of the organizations in our sample. Given the large numbers of institutions in the sample, we plot the coefficients for each organization with a 95% confidence interval in **Figures 7–10**.[29] In each figure, the vertical axis plots the effect. The horizontal axis enumerates the confidential institution identifiers. Recall that we express effects in terms of student standard deviation differences from the mean organization effect. An estimate of 0.10 therefore indicates that graduates from that organization produce leaning gains, on average, of 0.10 student standard deviations more than the average organization in our sample. We estimate that one standard deviation in teacher quality in this sample to be about 0.22 in math and 0.20 in ELA, so an institution effect of 0.10 represents about half a standard deviation in the teacher quality distribution.

[29] Point estimates and standard errors are shown in Appendix A.

Figure 7. Institution Effects (Math, Baseline Model)



Note: Sponsoring organization coefficients from baseline model of student achievement. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly higher than the state average; red estimates are statistically significantly lower than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level. The numbers on the horizontal axis represent a common identifier unique to this report to protect the confidentiality of each EPP.

Before discussing the estimates, we make one note about the standard for identifying exceptional institutions. In the figures that follow, the solid vertical line indicates the average of the institutions in the state. The vertical brackets indicate 95% confidence and the color indicates statistical significance. We use 95% confidence intervals because they are the standard in social science research, not because they should necessarily guide policymaking or practice. The confidence intervals cover the range of estimates we would expect to observe in 95% of similar samples of teachers. By definition, we would expect a program with state average performance to receive a statistically significant estimate 5% of the time. This is a *false positive*: we identify an average program as differing from the mean. The use of smaller confidence intervals raises the likelihood that extreme programs will be correctly identified as such, but also increases the probability of false positives. It is important to note that the choice of a threshold for accountability purposes should consider the consequences of labeling an EPP as high-performing or low-performing and the standard adopted in this report may be either too conservative or insufficiently conservative for any particular policy application.

The results in Figure 7 illustrate a few general features that are common to all of our estimates of average institutional effectiveness measures. First, we do observe a range of institution effects: the least effective program identified in this sample has an institution effect of about -0.11 and

the most effective program has an estimated effect of about 0.22. Taken at face value, this is a large difference: a difference of 0.33 standard deviations represents more than 75% the learning gains a typical student makes in a school year. Second, however, is the fact that confidence intervals for many programs cover a large range. Note that even for the high- and low-performing programs, we cannot rule out a wide range of actual average performance. Only 8 of the 37 institutions have point estimates that are statistically distinguishable from the mean institution. However, we observe a small number of teachers in many programs. This limits the precision with which we can estimate institution effects, particularly for smaller programs or those placing few candidates in in-state public schools. This problem has been widely noted in the literature on educator preparation programs in other states. Consequently, the precise ordering of institutions, particularly those in the middle of the distribution where differences are relatively small, are likely to fluctuate across samples.

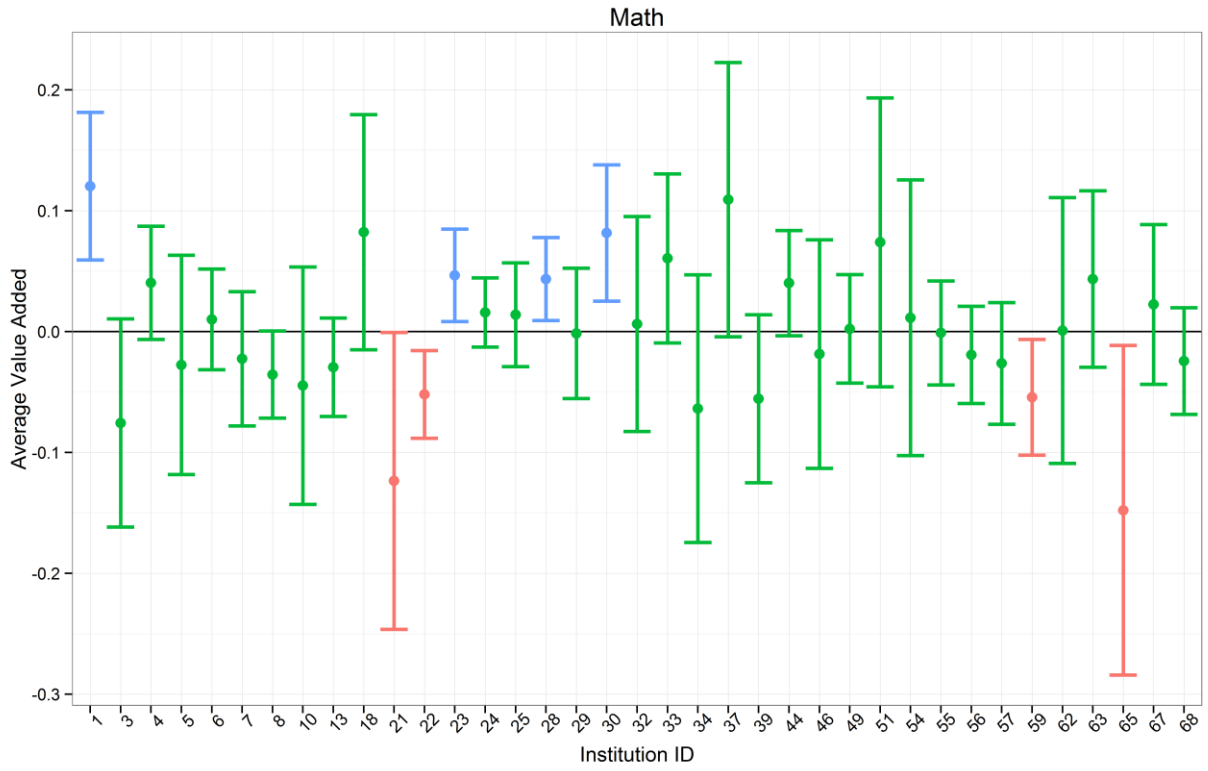
Although most institutions are not statistically significantly different than the mean institution, we do identify some outliers. Graduates of Programs 32 (-0.11), 22 (-0.06), 57 (-0.06), and 8 (-0.05) are all statistically significantly less effective than the mean institution. On the high end, completers from Programs 21 (0.22), 51 (0.18), 63 (0.09), and 33 (0.06) are all statistically significantly more effective than the mean institution. For comparison, the absolute differences between value added at these institutions and the state mean are all greater than the difference between a novice and fourth year teacher.[30] In Figure 8, we show the results when we add school fixed effects. Note that the school fixed effects models reduce the point estimates for the two highest-achieving programs under the baseline model. The coefficient for Program 21 falls from 0.22 to -0.12 and the coefficient for Program 51 falls from 0.18 to 0.07. Among those programs with statistically significantly negative impacts in the school fixed model, only Program 22 (-0.05) has a statistically significant effect in the baseline model as well. None of the programs with positive effects in the school fixed effects model also has a statistically significant effect in the baseline model, although the point estimate for Program 28 (baseline: 0.03, school fixed effects: 0.04) does not change substantially between models.

In other cases, however, institutions may be connected to particular school systems in ways that would cause the results of the two models to diverge. Some of the providers in our dataset are charter authorizers who primarily place their completers in their own schools. Others also disproportionately place students in charter schools. Research using the randomization of students that results from lotteries into these charter schools suggests that they are more effective in raising student achievement than traditional public schools (Abdulkadiroglu et al., 2011; Angrist et al., 2013). The baseline models may overstate the effectiveness of completers of these programs because they incorporate differences in the effectiveness of the schools. The problem is further exacerbated by the concentration of teachers from these programs in a small number of schools. Teachers who complete programs offered by charter authorizers may teach in schools that are staffed primarily by other teachers from their program. Charter schools may make different hiring decisions than traditional public schools and the within-school relationships may not generalize to other settings (Hoxby, 2002). More generally, these results highlight the fact that there will be greater modeling uncertainty for institutions that train teachers for very specific populations or schools. In these cases, decisions about how to model student achievement and

[30] We estimate differences in teacher effectiveness by experience levels using value added models with the covariates described in the methods section.

account for unobserved heterogeneity are likely to lead to larger changes in the estimated institution effects.

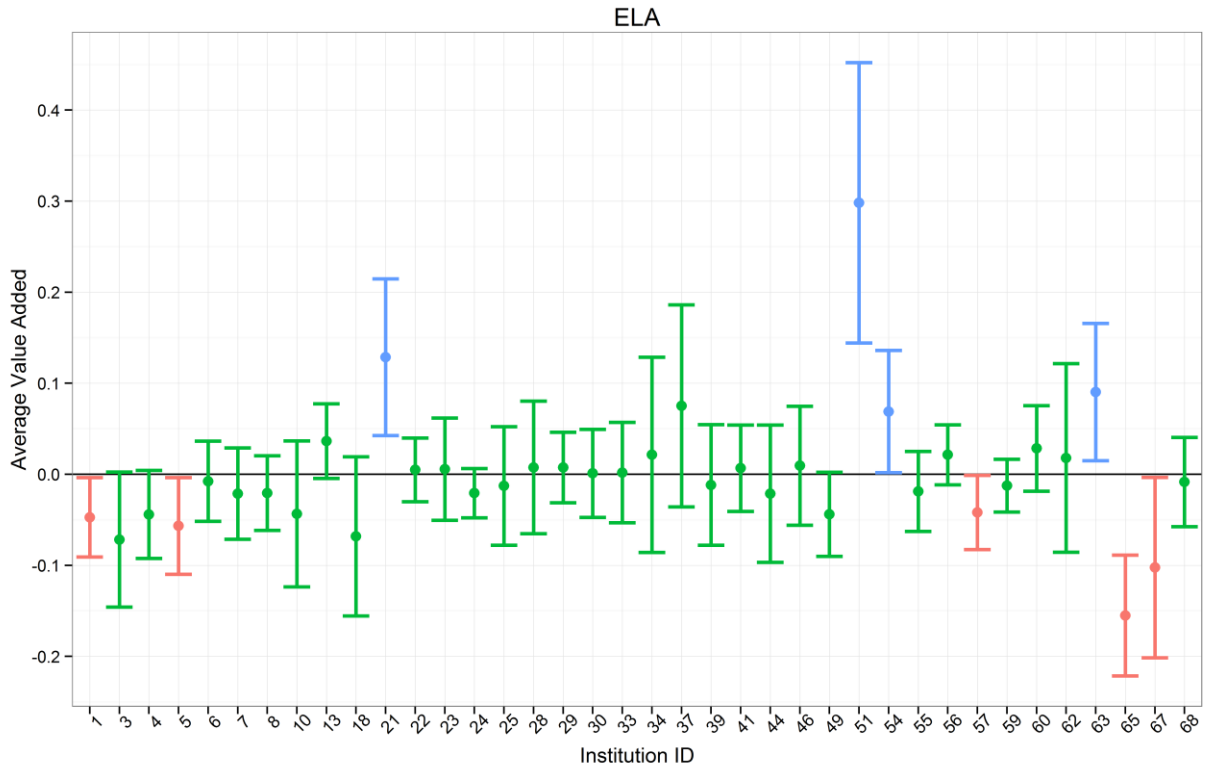
Figure 8. Institution Effects (Math, School Fixed Effects)



Note: Sponsoring organization coefficients from school fixed effects models of student achievement. Institution identifiers are displayed to the left of the estimates. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly higher than the state average; red estimates are statistically significantly lower than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level. The numbers on the horizontal axis represent a common identifier unique to this report to protect the confidentiality of each EPP.

We present comparable results for the ELA models in Figures 9 and 10. In the baseline models, we identify five organizations with estimated effects that are statistically significantly less effective than the mean institution. These are Program 65 (-0.16), 67 (-0.10), 5 (-0.06), 1 (-0.05), and 57(-0.04). Of these institutions, only Program 57 also has a statistically significantly negative effect in the baseline math model. At the other extreme, Programs 51 (0.30), 21 (0.13), 63 (0.09), and 54 (0.07) have statistically significantly positive estimates. The first three of these are also statistically significantly positive in the baseline math model as well. As is the case with the math results, differences between the exceptional programs

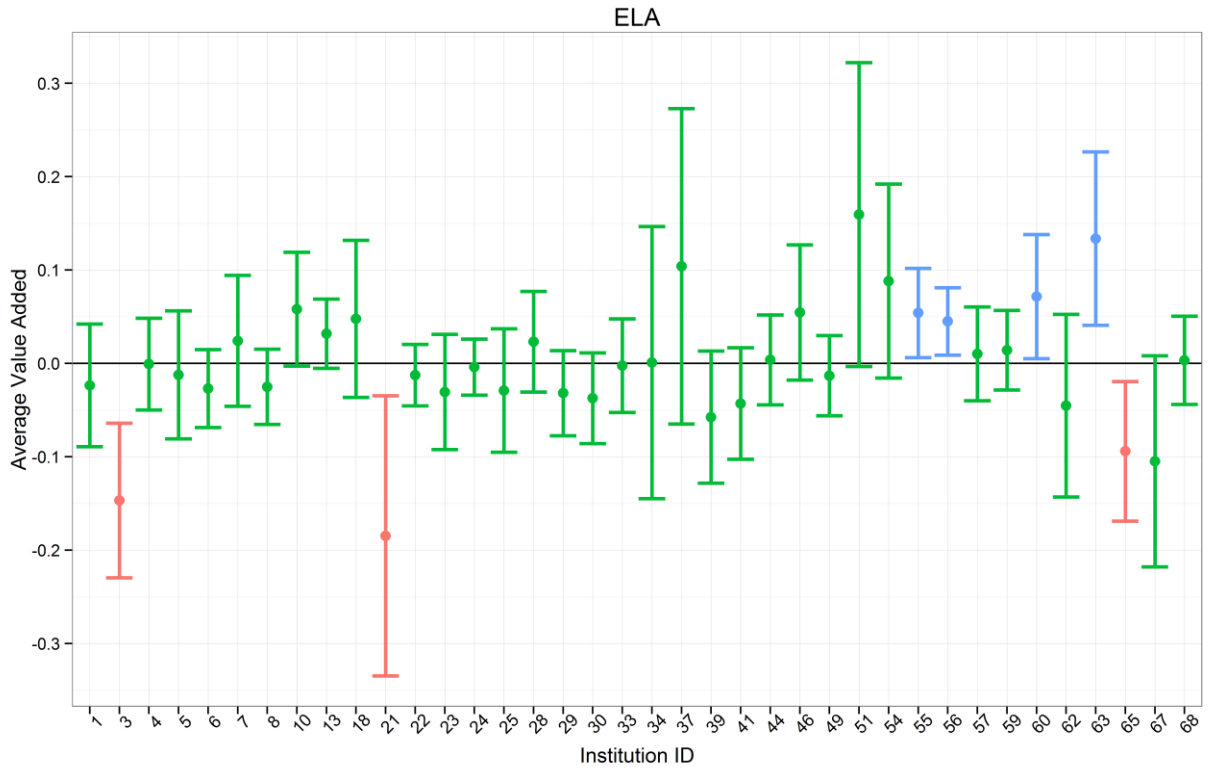
Figure 9. Institution Effects (ELA, Baseline Model)



Note: Sponsoring organization coefficients from baseline model of student achievement. Institution identifiers are displayed to the left of the estimates. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly higher than the state average; red estimates are statistically significantly lower than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level. The numbers on the horizontal axis represent a common identifier unique to this report to protect the confidentiality of each EPP.

In Figure 10, we add the school fixed effects. We observe a similar pattern for the two outliers as with the math achievement analysis. The coefficients for both Program 21 (0.13 to -0.18) and Program 51 (0.30 to 0.16) drop significantly. The latter institution is not statistically significantly different from the mean in the school fixed effects model. In this specification, Programs 21 (-0.18), 3 (-0.15), and 65 (-0.09) are statistically significantly negative. Program 65 is the only institution that appears with a statistically significantly negative estimate in both models, although point estimates for Program 67 are similar across models. On the positive side, Programs 63 (0.13), 60 (0.07), 55 (0.05), and 56 (0.05) are all statistically significant; only Program 63 appears in both lists.

Figure 10. Institution Effects (ELA, School Fixed Effects)



Note: Sponsoring organization coefficients from school fixed effects models of student achievement. Institution identifiers are displayed to the left of the estimates. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly higher than the state average; red estimates are statistically significantly lower than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level. The numbers on the horizontal axis represent a common identifier unique to this report to protect the confidentiality of each EPP.

The value-added models identify a few common institutions at the extremes across subjects and model specifications, but as Figures 7–10 indicate, there is substantial disagreement in the precise ordering of programs. In order to understand their overall agreement, we estimate the correlation in the estimates of individual institutions across models. The correlation in the institutional effectiveness measures between the baseline math and baseline ELA model is about 0.62. This is similar to the correlation between estimates of math and ELA value added for individual teachers (Goldhaber et al., 2013a). There is less agreement among the different modeling choices. For math, the school fixed effects and baseline value-added measures correlate at 0.11 and the ELA measures correlate at 0.47. Similar levels of modeling disagreement have been found in analyses of Texas programs (von Hippel et al., 2016). These lower correlations highlight the importance of considering school context, although as we note above, the small number of teachers working in many of the schools in our sample generates much less precisely estimated estimates when including school fixed effects.

The student achievement analysis suggests that there are differences in value added across institutions in Massachusetts, but this analysis does not give a sense of the relative magnitude of these estimates compared to the overall variation in teacher effectiveness. We therefore

decompose the variation in estimated teacher effects in **Table 8.** In columns 1-2 and 5-6 of Table 8, we calculate the standard deviation of institution and within-institution teacher effects (i.e., the variation in value added for teachers who attended the same institution) in Math (columns 1-2) and ELA (columns 5-6). In columns 3-4 and 7-8 of Table 8, we also consider specific programs within institutions and calculate the standard deviation of institution effects, within-institution program effects, and within-program teacher effects. Because there is not a single preferred method for this type of variance decomposition in the empirical literature, we use two approaches to estimate the variance in teacher effectiveness across programs and institutions; odd columns present estimates from random effects models, while even columns present estimates from fixed effects models as implemented in Koedel et al. (2016). We present both sets of results because they suggest slightly different conclusions about the variation between and within institutions.[31]

Table 8. Variation in Institution, Program, and Teacher Effects on Student Achievement

Component	Estimated Component Standard Deviation							
	Math				English/Language Arts			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Institution	0.080	0.041	0.047	0.041	0.035	0.036	0.000	0.036
Program			0.097	0.033			0.060	0.033
Teacher	0.203	0.173	0.189	0.170	0.197	0.171	0.190	0.167

Note: The estimated variance components are derived from the random effects model described in the text. For the results in odd-numbered columns, we first estimated models with controls for teacher experience, school demographics, and teacher, grade, and year effects. We then estimated the variance components from the residuals (combined residuals and teacher effects) produced by these models. Each of the variance components models included both teacher and teacher-year effects. For the results in even-numbered columns, we used the change in the R^2 measure from the addition of the relevant effects as described by Koedel et al. (2015b). We estimated these models using programs with at least 10 completers in the student achievement sample.

There are three broad conclusions from this exercise, each corresponding to a row in Table 8. The first row of Table 8 shows that the standard deviation of institution effects is about 0.041 to 0.080 standard deviations in student math achievement and about 0.035 standard deviations in ELA student achievement. One standard deviation in the institutional effectiveness distribution therefore corresponds to a difference of about 4-8 weeks of student learning (Bloom et al., 2008). Put another way, about 5–13% of the variance of math value added and about 3–4% of the variance of ELA value added can be explained by the average value added from a teacher's institution.[32] These results are within the range of those estimated in other states, although the variation in math value added is on the higher end of estimates from other states and indicates that institutions in Massachusetts are more variable in the effectiveness of their graduates than elsewhere.

[31] In addition to the methods presented in Table 8, we also employed two additional approaches adopted in the literature. First, we computed the variance of institution effects as the covariance between two random sets of programs in the same institution. We then estimated the covariance between teacher effects of two random teachers who completed the same program and deducted the variance of institution effects to calculate the program effects variance. This produces similar results as the random effects method. Second, we used the F-statistic approach described by Koedel et al. (2015b). This method produces results similar to the R^2 method.

[32] This is estimated as the square of the proportion of the institution standard deviation to the combined teacher standard deviation.

The second row of Table 8 shows that the variation in value added across programs within an institution is comparable to—and in some cases, larger than—the differences across institutions. In other words, the expected difference in value added for two teachers selected from different programs within the same institution is at least as large as the expected difference in value added for teachers selected from two different institutions. Given the importance of individual programs for explaining teacher value added, knowing both a teacher's program and institution provides more predictive power than knowing the institution alone. We quantify this explanatory power by estimating the share of the overall variation in teacher effectiveness – inclusive of the institution, program, and unobserved teacher factors – that is explain by the institutions and programs. Adding program effectiveness, the proportion of value added explained by preparation background increases from 5-13% to 10-25% in math and from 3-4% to about 9% in ELA.[33]

That said, the final row of Table 8 reinforces a consistent finding from the empirical literature on institution and program effects; namely, that there is far more variation in teacher value added within institutions and programs than across institutions and programs. Specifically, the expected difference in teacher value added for two teachers selected from within the same institution or program is 2-5 times larger than the expected difference in average value added between two different programs or institutions.

Preparation Pathways and Summative Performance Ratings (RQ 2)

The value-added analyses provide evidence on the effectiveness of Massachusetts teachers only for a limited number of grades and subjects. We therefore assess preparation program effects on teacher performance on Massachusetts' summative performance rating (research question #2). The summative performance rating is a final assessment of teacher performance under each of the four teaching standards. As we describe above, districts establish their own procedures for determining the performance ratings and for aggregating teacher performance on each of the standards into a final performance rating. Because these rules tend to take the form of cut points for establishing teacher proficiency, there is more variation in the individual standards scores than on the final summative performance rating. In this analysis, we therefore use averages of the four standards as our outcome of interest.[34]

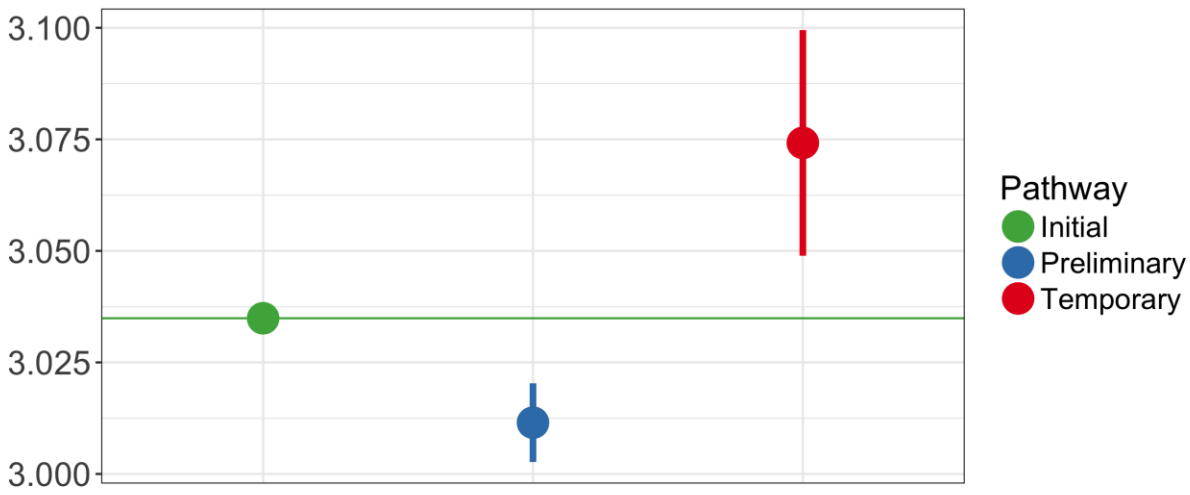
Licensure Pathways and Summative Performance Ratings

We begin by examining differences in educator effectiveness across licensure pathways in **Figure 11**. Figure 11 plots the expected summative rating by licensure pathway. These results rely on the models with controls for school characteristics, but the patterns hold in each of the models we estimate. The mean rating in the sample is just above 3.0, which corresponds to a proficient rating. Teachers entering with preliminary licenses earn the lowest summative ratings, and these differences are statistically significant. At the other end, teachers with temporary licenses earn ratings that are about 0.04 points higher than those with initial licenses.

[33] As above, this share is estimated as the proportion of the combined variance in teacher quality in Table 8 explained by the institution and program factors.

[34] We have also explored using the final summative performance rating and the first principal component of the four standard scores. These analyses produced similar results.

Figure 11. Summative Ratings by Licensure Pathway



Note: Preliminary (temporary) license indicates that a teacher's first Massachusetts license is of the preliminary (temporary) type. Teachers who entered with an initial license comprise the omitted group. The estimates are derived from regressions of the mean rating across the four professional standards on indicators for initial program type. All models include controls for teacher experience, school demographics, and year effects. We estimate predicted values using the mean characteristics. Standard errors account for clustering at the teacher level.

We quantify these differences in **Table 9**. The results in Table 9 are based on regressions of the average teacher rating on the four standards on institution indicators, teacher experience, and school demographic characteristics. We translate the ratings into a quantitative variable by ranking the rating categories on a four-point scale. The pathway effects therefore represent differences in the average points awarded for each licensure pathway. As before, teachers with initial licenses are the reference group in these regressions and point estimates indicate differences in average ratings with the indicated license type. Overall, the summative performance data provide a clearer picture of teacher effectiveness differences than in the value-added results. The estimates are relatively stable across the models and indicate that teachers with preliminary licenses earn ratings about 0.03 points lower than teachers with initial licenses. The direction of the estimates is consistent with the differences we observe with math value added. Teachers with temporary licenses earn ratings of about 0.03–0.04 points higher than teachers with initial licenses. Both of these differences correspond to about 0.10 standard deviations in the average ratings, or slightly less than the difference in mean ratings between a novice teacher and a teacher with 1 year of experience.

Table 9. Average Differences in Summative Ratings Relative to Teachers with Initial Licenses

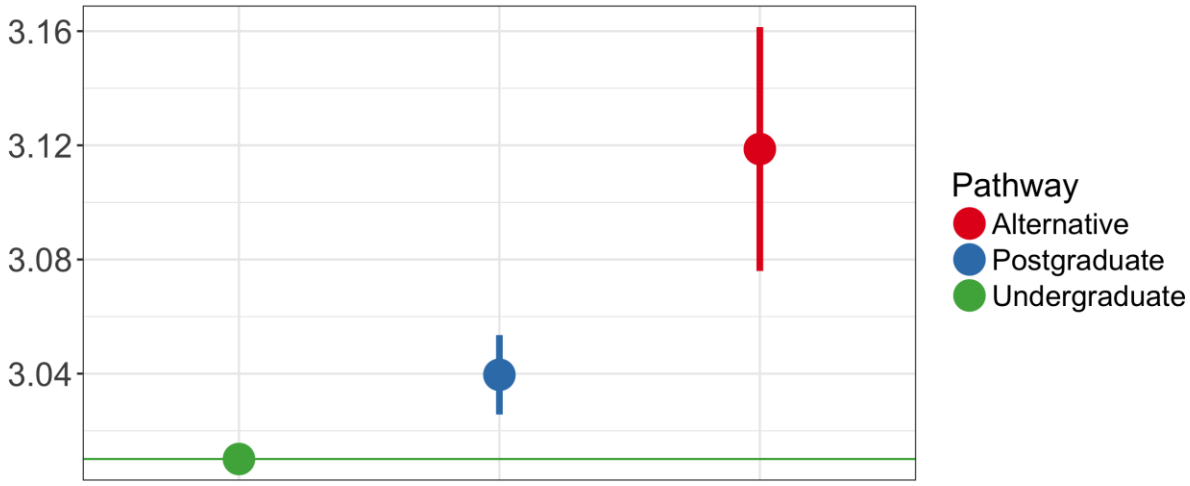
Licensure Pathway	Baseline Model (1)	District Fixed Effects (2)	School Fixed Effects (3)
Preliminary License	-0.023*** (0.004)	-0.026*** (0.004)	-0.022*** (0.004)
Temporary License	0.039*** (0.013)	0.028** (0.012)	0.034** (0.013)

Note: Preliminary (temporary) license indicates that a teacher's first Massachusetts license is of the preliminary (temporary) type. Teachers who entered with an initial license comprise the omitted group. The estimates are derived from regressions of the mean rating across the four professional standards on indicators for initial program type. All models include controls for teacher experience, school demographics, and year effects. The models in columns 2 and 3 control for school and district fixed effects. Standard errors (in parentheses) account for clustering at the teacher level. N = 24,602. *p < 0.10, **p < 0.05, ***p < 0.01.

Program Type and Summative Performance Ratings

In **Figure 12**, we restrict our attention to the program completer data and estimate differences in assessed effectiveness across program types.

Figure 12. Summative Ratings by Program Pathway



Note: Postgraduate (alternative) program indicates the type of Massachusetts program a teacher attended. Teachers who attended an undergraduate program comprise the omitted group. The estimates are derived from regressions of the mean rating across the four professional standards on indicators for initial program type. All models include controls for teacher experience, school demographics, and year effects. We estimate predicted values using the mean characteristics. Standard errors account for clustering at the teacher level.
N = 12,381. *p < 0.10, **p < 0.05, ***p < 0.01.

We plot the predicted ratings by program pathway. In contrast to the findings on licensure pathway, we find that teachers from alternative programs, who generally enter with a preliminary license, earn the highest summative ratings. The difference between undergraduate and alternative programs corresponds to about one-third of a standard deviation in mean ratings, or approximately the difference between a novice teacher and a teacher with five years of experience. This finding is in contrast with the value-added results, where we found little

consistent evidence that teachers from alternative programs were more effective than those from undergraduate programs. Teachers from traditional programs have more similar ratings, but those from postgraduate programs perform somewhat higher than those from undergraduate programs. In this case, the difference between undergraduate and postgraduate programs corresponds to a difference of about 0.10 standard deviations or about the difference between a novice and second year teacher

In **Table 10**, we test the sensitivity of these results to alternative specifications. As is the case with the licensure analysis, estimates using summative ratings are less sensitive to model specification than the corresponding analyses of teacher value added. We find evidence consistent with the ELA value-added analysis that teachers from postgraduate programs were more effective than teachers from undergraduate programs. In the summative performance data, we estimate average differences of about 0.02–0.03 points. The translation of these estimates into standard deviations in the distribution of ratings provides similar magnitudes as those found in the ELA analysis. The estimates for teachers from alternative programs suggest that they earn ratings that are 0.08–0.10 points higher, on average, than teachers from undergraduate programs.

Table 10. Average Differences in Summative Ratings Relative to Teachers from Undergraduate Programs

Program Pathway	Baseline Model (1)	District Fixed Effects (2)	School Fixed Effects (3)
Postgraduate Program	0.030*** (0.007)	0.024*** (0.006)	0.023*** (0.007)
Alternative Program	0.109*** (0.022)	0.078*** (0.020)	0.080*** (0.020)

Note: Postgraduate (alternative) program indicates the type of Massachusetts program a teacher attended. Teachers who attended an undergraduate program comprise the omitted group. The estimates are derived from regressions of the mean rating across the four professional standards on indicators for initial program type. All models include controls for teacher experience, school demographics, and year effects. The models in columns 2 and 3 control for school and district fixed effects. Standard errors (in parentheses) account for clustering at the teacher level.
N = 12,381. *p < 0.10, **p < 0.05, ***p < 0.01.

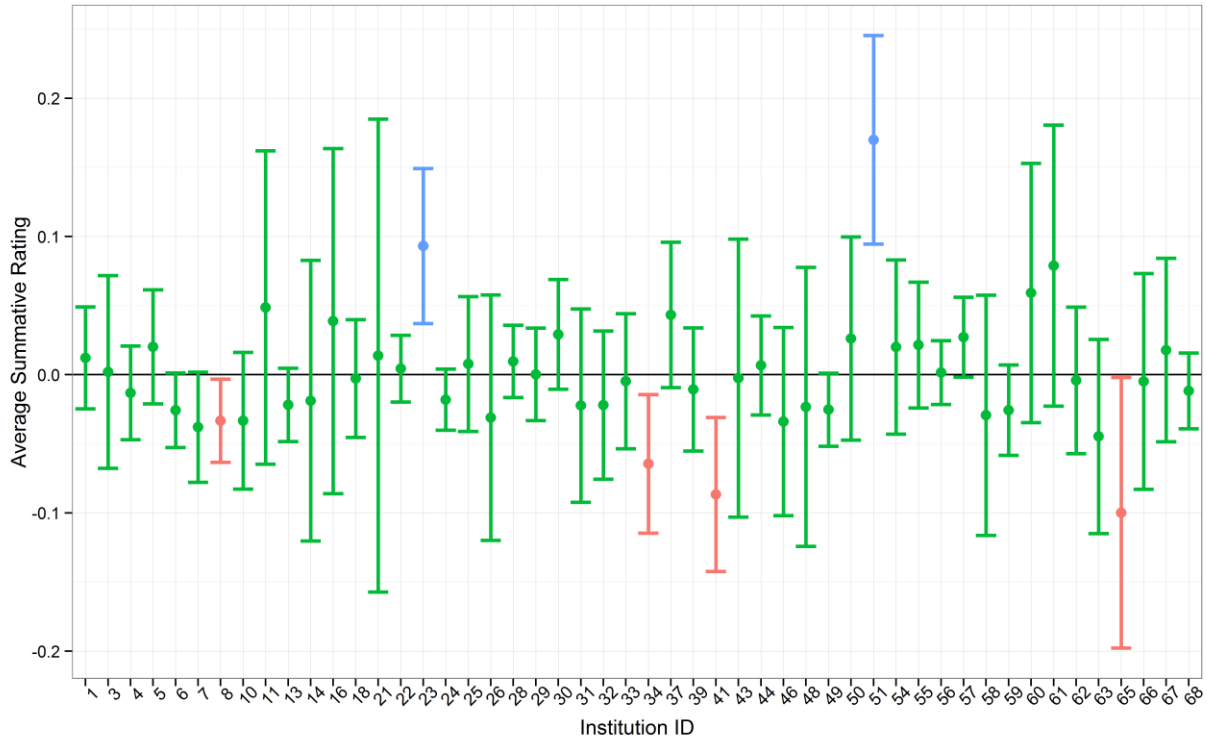
Programs, Institutions, and Summative Performance Ratings

We now turn to an analysis of individual institutions using the summative performance data. As with the student achievement data, we plot institution effects on the vertical axis. The horizontal axis enumerates the confidential institution identifiers. We estimate models similar to the pathway models above and the contrasts are measured in points on the rating scale. An estimate of 1.0 indicates that an institution's graduates score, on average, a full point higher on the summative rating scale than the mean program in the state. One standard deviation in the ratings is about 0.30 points on the summative rating scale.

The same caveats apply to the estimates of average institutional ratings as to the results on value added. In many cases, we observe only a small number of teachers per institution and some of the estimated averages are imprecisely estimated. Estimating these same models on different cohorts of teachers would likely lead to considerable resorting, particularly among institutions in the middle of the distribution with similar average ratings. We use the same confidence intervals in these estimates and denote statistical significance using the color of the bars. However, as

before, 95% confidence intervals may not be an appropriate choice for any particular policy application.

Figure 13. Institution Effects (Summative Performance Ratings)



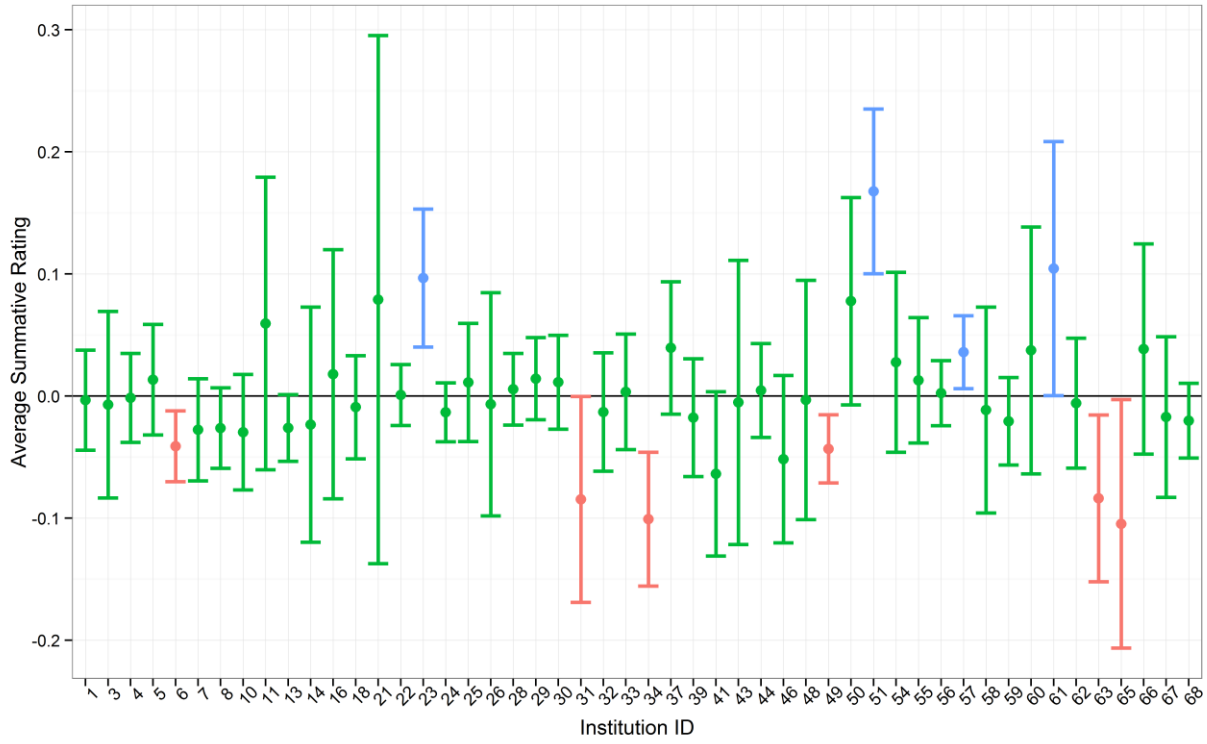
Note: Sponsoring organization coefficients from district fixed effects models of summative performance ratings. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly higher than the state average; red estimates are statistically significantly lower than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level.

We plot the results from a model with district fixed effects in **Figure 13**. These models use only within-district variation in summative ratings to avoid conflating institution effects and district-level variation in the implementation of the educator evaluation framework. We identify four programs at the low end of the distribution of institution effects that differ statistically significantly from the mean: Programs 65 (-0.10), 41 (-0.09), 34 (-0.06), and 8 (-0.03). At the high end, only Programs 51 (0.17) and 23 (0.09) are statistically significant.

In **Figure 14**, we replaced the district fixed effects with school fixed effects. As with the student achievement models, introducing school fixed effects reduces the precision of our estimates because we are comparing each teacher only to others in the same schools. Nonetheless, the top and bottom programs are similar between the two models. Estimates for Programs 51 (0.17) and 23 (0.10) remain statistically significantly above the mean in the school fixed effects model. Teachers from Program 61 also have statistically significantly more favorable ratings. Among the bottom programs, teachers from Programs 65 (-0.10) and 34 (-0.10) have statistically significantly weaker evaluations in both models. In the school fixed effects models, they are additionally joined by Programs 31 (-0.09), 63 (-0.08), 49 (-0.04), and 6 (-0.04). In the case of

Programs 6 and 49, although the statistical significance changes, the point estimates are similar to those in the district fixed effects models.

Figure 14. Institution Effects (Summative Performance Ratings, School Fixed Effects)



Note: Sponsoring organization coefficients from school fixed effects models of summative performance ratings. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly higher than the state average; red estimates are statistically significantly lower than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level.

The institution effects we estimated in Figures 13 and 14 are jointly statistically significant, which suggests that there is variation in teacher effectiveness ratings across programs. As with the value added measures, we estimate this variability directly in **Table 11** by decomposing teacher ratings into components explained by institutions, programs, and teachers. In columns (1) and (2), we first estimate the variability of institutions. These measures describe the dispersion between mean teacher ratings at each institution from the state mean. We find that a standard deviation in the institution effects is about 0.01 to 0.03, depending on the procedure. The standard deviation of overall teacher effectiveness, inclusive of both institution effects and the unobserved teacher component, is about 0.165 to 0.251 standard deviations. Using either procedure, we find that the institution indicators therefore explain about 1% of the variance in summative performance ratings.[35]

[35] As is the case with the value-added estimates, it is necessary to square both terms in order to compute the variance.

Table 11. Variation in Institution and Program Effects on Summative Performance Ratings

Component	Estimated Component Standard Deviation			
	(1)	(2)	(3)	(4)
Institution	0.014	0.030	0.013	0.030
Program			0.019	0.049
Teacher	0.165	0.249	0.164	0.244

Note: The estimated variance components are derived from the random effects model described in the text. For the results in odd-numbered columns, we first estimated models with controls for teacher experience, school demographics, and teacher, grade, and year effects. We then estimated the variance components from the residuals (combined residuals and teacher effects) produced by these models. Each of the variance components models included both teacher and teacher-year effects. For the results in even-numbered columns, we used the change in the R^2 measure from the addition of the relevant effects as described by Koedel et al. (2015b). We estimated these models using programs with at least 10 completers in the student achievement sample.

In columns (3) and (4), we add program indicators. The standard deviation in the program effects describes the dispersion between mean ratings for each program and the overall mean rating within the institution. As is the case with value-added estimates, we find that the contribution of programs is empirically important. Depending on the model, we estimate that one standard deviation in program effects is about 0.02–0.05, or about 1–2% of the teacher-level variance in summative ratings. In both cases, the variation across programs is at least as large as the variation across institutions. In other words, two programs selected at random from within an institution are likely to have a greater difference in ratings than two randomly selected institutions. Jointly, programs and institutions explain about 3–6% of the variance in summative ratings. This is lower than we observe for the value-added measures, but is consistent with estimates of institution effects using observational measures of teacher quality from Tennessee (Ronfeldt & Campbell, 2016).

Results: Teacher Preparation and Retention

Licensure Pathways and Teacher Attrition

Teachers from difference licensure pathways have considerably different prior experiences in teaching. Teacher attrition is highest during the first few years in the classroom, so small differences in prior experiences may generate substantial differences in the likelihood that teachers remain in the classroom. Because of the increased attrition during the early portion of teachers' careers, we first focus on differences in retention by year of experience before moving to more parsimonious summaries of differences in average attrition rates. As shown in Table 2, there are some differences in school characteristics across the licensure pathways. In order to account for differences in the school context in which teachers work, we base these probabilities on the estimated results from the models described in the Methods section.[36]

The overall attrition rates during the first few years in the classroom are roughly consistent with national survey data (Goldring et al., 2014). For teachers with initial licenses, the rate of departure is 9.5% during the first year but falls to 5.6% by the fourth year. Both of the other

[36] In order to capture variability across pathways in the timing of attrition, we interact pathway indicators with the experience measures. We drop these terms when we estimate differences in the average probability of attrition.

licensure pathways have higher rates of attrition, especially initially. About 12.5% of teachers with preliminary licenses and 16% of teachers with temporary licenses leave after the first year of teaching. As is the case with teachers entering with initial licenses, the probability of departure declines thereafter. Even by the fourth year of teaching, however, teachers from both of these pathways leave at higher rates than teachers with initial licenses.

Table 12. Licensure Pathways and Average Teacher Attrition

		Estimated Attrition Rate		
Licensure Pathway	**Experience**	**Baseline Model (1)**	**School Random Effects (2)**	**School-Year Random Effects (3)**
Initial License	0	9.5	8.3	9.1
	1	8.2	7.9	8.3
	2	7.0	7.2	7.0
	3	5.6	6.0	5.9
Preliminary License	0	12.5	10.6	11.6
	1	10.0	9.4	9.7
	2	7.3	7.2	7.1
	3	7.0	7.5	7.3
Temporary License	0	16.0	13.6	14.4
	1	15.3	13.7	14.2
	2	9.1	8.6	8.6
	3	8.5	7.8	7.8

Note: License pathway indicates the first type of Massachusetts teaching license earned. Estimated predicted probabilities (percentages) are derived from the probit models of teacher exit described in the text with the addition of interactions between license type and tenure. For all estimates, we fix all school characteristics at the sample mean and vary license type and tenure as indicated in the table. All models include controls for teacher experience, school demographics, and year effects. The models in columns 2 and 3 control for school unobservables using the procedures described in the text. *N* = 33,564.

The results from the more formal attrition models, which we present in **Table 13**, provide the same ordering of licensure pathways. We display the results as the average marginal effects so that the coefficients provide the difference in the probability of leaving Massachusetts public schools. As we now switch to showing results in terms of the probability of leaving Massachusetts schools, an estimate of 0.01 therefore suggests that teachers from the given pathway have a probability of leaving the profession that is 1 percentage point higher than teachers with an initial license (the omitted category). The results in column 1indicate that, all else being equal, teachers who enter the profession with preliminary licenses are about 2 percentage points more likely to leave the teaching profession each year than teachers with initial licenses. Teachers with temporary licenses are about 4–5 percentage points more likely to leave Massachusetts public schools. These estimates are robust to the various modeling choices we make in Table 15, including methods that rely on comparisons within schools. About 9% of teachers in this sample depart the public school system each year, so the differences across pathways are a substantial proportion of the average attrition rate.

Table 13. Average Differences in Attrition Relative to Teachers with Initial Licenses

	Average Marginal Effects		
Licensure Pathway	**Baseline Model** **(1)**	**School Random Effects** **(2)**	**School-Year Random Effects** **(3)**
Preliminary License	0.019*** (0.003)	0.017*** (0.004)	0.016*** (0.004)
Temporary License	0.049*** (0.008)	0.042*** (0.009)	0.041*** (0.010)

Note: Preliminary (temporary) license indicates that a teacher's first Massachusetts license is of the preliminary (temporary) type. Teachers who entered with an initial license comprise the omitted group. Estimated marginal effects are derived from the probit models of teacher exit described in the text. All models include controls for teacher experience, school demographics, and year effects. The models in columns 2 and 3 control for school unobservables using the procedures described in the text. Standard errors (in parentheses) are computed using the delta method and account for clustering at the teacher level. N = 33,564. *p < 0.10, **p < 0.05, ***p < 0.01.

Program Type and Teacher Attrition

We next consider attrition patterns among teachers in the program completer sample. Before turning to the estimates for individual institutions, we present results by program type. The estimated attrition rates by year in **Table 14** are more similar across program type than the comparable estimates by licensure type in Table 12. About 7.1% of teachers from an undergraduate program leave after the first year. Estimates for teachers from postgraduate programs are slightly higher at about 7.5%. In both cases, the probability of attrition subsequently falls so that, by the fourth year, it is about 3.6% and 5.0%, respectively. The attrition rates are higher for teachers from alternative programs. Nearly 10% of teachers from these programs leave after the first year. The probability of attrition remains higher throughout our sample: by the fourth year, attrition rates are still about 3-5 percentage points higher than the other program types. The difference we observe between traditional teacher preparation programs and alternative programs is consistent with analyses of national surveys and state administrative data (Boyd et al., 2006; Redding & Smith, 2016).

Table 14. Program Pathways and Average Teacher Attrition

		Estimated Attrition Rate		
Program Pathway	**Experience**	**Baseline Model (1)**	**School Random Effects (2)**	**School-Year Random Effects (3)**
Undergraduate Program	0	7.1	5.5	6.2
	1	6.2	5.5	5.8
	2	5.4	5.3	5.3
	3	3.6	3.8	3.5
Postgraduate Program	0	7.5	6.3	7.3
	1	7.6	6.9	7.7
	2	5.9	6.1	6.2
	3	5.0	5.6	5.2
Alternative Program	0	9.8	7.1	8.3
	1	6.3	5.4	6.4
	2	7.5	6.7	7.2
	3	8.6	8.5	7.6

Note: Program pathway indicates the first type of Massachusetts teacher preparation program completed. Estimated predicted probabilities (percentages) are derived from the probit models of teacher exit described in the text with the addition of interactions between program type and tenure (experience). For all estimates, we fix all school characteristics at the sample mean and vary program type and tenure as indicated in the table. All models include controls for teacher experience, school demographics, and year effects. The models in columns 2 and 3 control for school unobservables using the procedures described in the text. Standard errors (in parentheses) are computed using the delta method and account for clustering at the teacher level. N = 17,661.

We summarize these more succinctly in **Table 15** as the average difference in attrition rates between teachers from undergraduate programs and teachers from the other program pathways. Note that we again switch to probability units to characterize the effects as changes in the likelihood of attrition. Teachers from postgraduate programs leave Massachusetts public schools at rate of about 1–2 percentage points per year higher than undergraduate completers. These results are statistically significant in models that control for school effects and marginally insignificant in the baseline model. However, the results for completers of alternative programs are less robust to model specification. Although the point estimates are between 1–2 percentage points in all models, the results are smaller and not statistically significant when controlling for school effects in column (2).

Table 15. Average Differences in Attrition Relative to Teachers from Undergraduate Programs

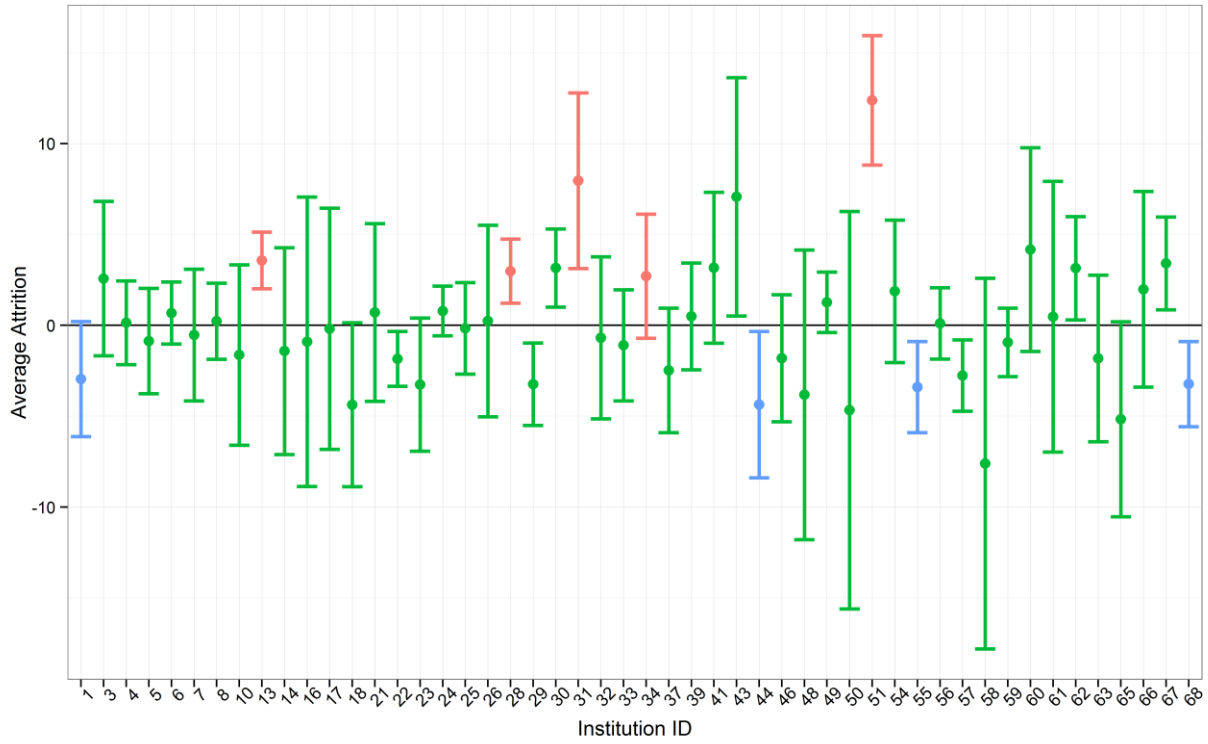
	Average Marginal Effects		
Program Pathway	**Baseline Model** **(1)**	**School Random Effects** **(2)**	**School-Year Random Effects** **(3)**
Postgraduate Program	0.008* (0.005)	0.012** (0.005)	0.015*** (0.005)
Alternative Program	0.019*** (0.008)	0.014 (0.010)	0.019* (0.010)

Note: Postbaccalaureate (alternative) program is the type of Massachusetts program a teacher attended. Teachers who attended a baccalaureate program comprise the omitted group. Estimated marginal effects are derived from the probit models of teacher exit described in the text. All models include controls for teacher experience, school demographics, and year effects. The models in columns 2 and 3 control for school unobservables using the procedures described in the text. Standard errors (in parentheses) are computed using the delta method and account for clustering at the teacher level. N = 17,661. *p < 0.10, **p < 0.05, ***p < 0.01.

We present similar results for estimates of the individual organization attrition rates. We display the regression estimates of differences in the annual attrition rates in **Figures 15** and **16**. We first estimate the regression models described in the methods section and then compute average marginal effects by institution centered around 0. The estimates represent differences in the annual attrition rates from the state average. Lower attrition rates are denoted by negative point estimates, so, unlike the figures for student achievement and the summative performance indicators, lower values correspond to better outcomes. Most of the institution effects are located within 5 percentage points of the state mean. There are a few outliers with high attrition rates: Programs 51 (+12), 31 (+8), and 40 (+7). Each of these is statistically significantly higher than the mean attrition effect. A few other institutions with smaller attrition effects are also statistically significant, including Programs 13 (+4), 67 (+3), 30 (+3), 62 (+3), and 28 (+3). At the other extreme, Programs 44 (-4), 55 (-3), 29 (-3), 68 (-3), 57 (-3), and 22 (-2) all have attrition rates that are statistically significantly lower than the average institution.

The estimates of individual institution attrition rates appear to be consistent with some general empirical findings in the research literature. Specifically, the institutions with high attrition rates include a number of more selective institutions. Teachers with higher college entrance examination scores or who attended more selective undergraduate institutions are more likely to leave the teaching profession (Guarino et al., 2006; Lankford et al., 2002; Podgursky et al., 2004). Selective institutions, which increasingly draw upon a national applicant pool, may have students with weaker connections to Massachusetts. Their graduates may also face greater wage offers outside education.

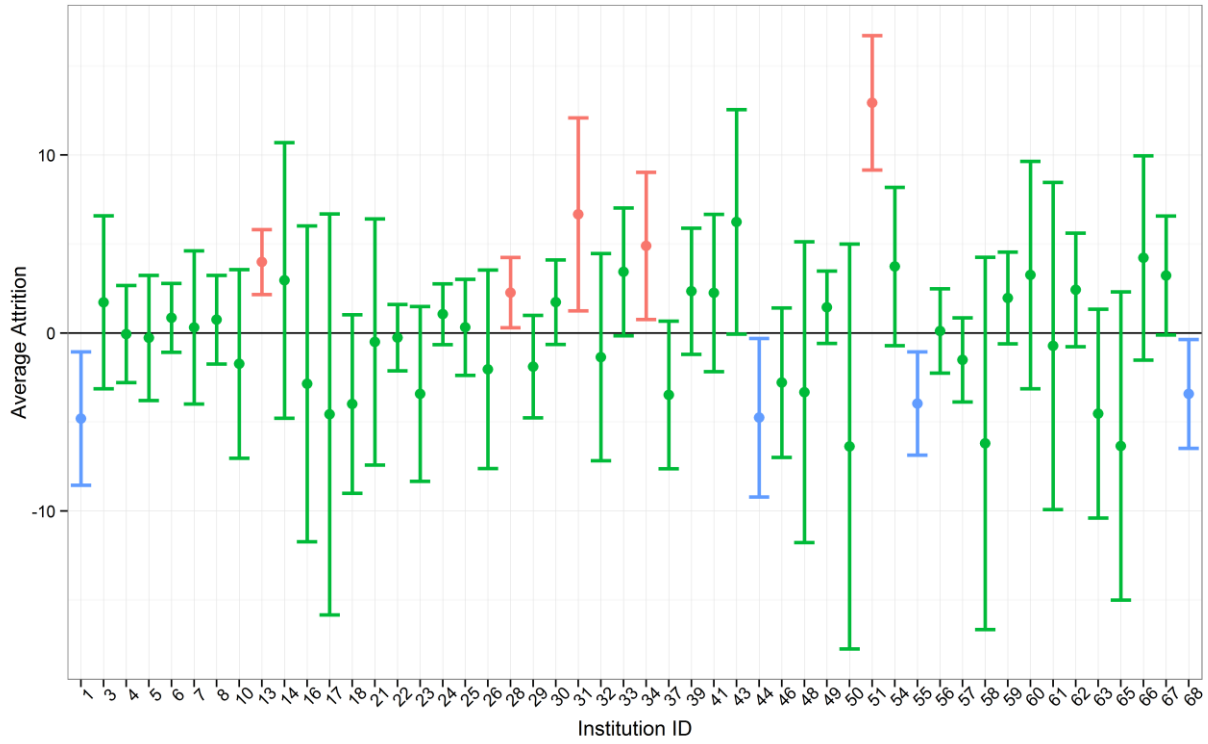
Figure 15. Attrition Effects



Note: Sponsoring organization coefficients from baseline models of teacher attrition. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly lower than the state average; red estimates are statistically significantly higher than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level.

The results in Figure 16 provide some adjustment for unobservable school characteristics using the correlated random effects approach described in the methods section. The two models generally agree on the ranking of individual institutions—the rank order correlation is 0.85—although the results in Figure 16 are somewhat less precisely estimated than the baseline estimates in Figure 15. The similarity in the two sets of estimates suggests that controlling for school context beyond the inclusion of student demographic characteristics is less important in the attrition models than in the summative assessment and student achievement analyses.

Figure 16. Attrition Effects (School Correlated Random Effects)



Note: Sponsoring organization coefficients from school correlated random effects models of teacher attrition. Circles indicate point estimates and solid lines indicate 95% confidence intervals. Blue estimates are statistically significantly lower than the state average; red estimates are statistically significantly higher than the state average. The organization indicators are expressed as deviations from the mean organization effect. Standard errors are clustered at the teacher level.

Discussion

Massachusetts licenses new teachers through three main channels. The most common route, initial licensure, is for teachers who have completed a course of study from an accredited institution and all of the state licensure requirements. Teachers with preliminary licenses have satisfied state testing requirements but have not completed an approved education program. This route is common among those attending alternative programs in the state. We found some evidence that teachers who obtained the preliminary license, including candidates in alternative programs and other individuals seeking a teaching position, as their first teaching license were less effective than those entering with the initial license. In most models, we found weaker student achievement gains in math. Teachers with preliminary licenses also earned lower summative ratings. These empirical findings appear to have more to do with prior teaching experience or preparation than the quality of alternative programs in Massachusetts. When we restricted the sample to program completers, teachers from alternative programs earn higher summative ratings than teachers from traditional programs and we found no evidence that they are less effective in the classroom.

The temporary license is designed for out-of-state teachers. These teachers have not completed state testing requirements but have completed an educator preparation program and 3 years of classroom teaching in another state. In a recent analyses of teachers in North Carolina public schools, Bastian and Henry (2015) found that teachers crossing state lines were less effective in the classroom. We find little evidence of such a phenomenon in Massachusetts. Teachers with temporary licenses are not less effective at raising student test scores and perform better on summative performance assessments. They do, however, have higher turnover rates, a finding consistent with the evidence from North Carolina.

Among those teachers earning initial licenses after training in Massachusetts, there is a great deal of diversity in preparation experiences. In-state providers include both traditional programs housed in colleges and universities and 16 providers that offer alternative routes into the profession. We find that differences in the average effectiveness of teachers across institutions explains about 3–4% of the variation in educator effectiveness in ELA and 5–13% of the variance in math. The variability among providers is toward the higher end of comparable analyses in other states. There may therefore be greater benefits to program improvement in Massachusetts than elsewhere. Similarly, the greater variation among institutions suggests that there may be more room to learn about effective practices from other in-state providers through technical assistance or other means of sharing best practices.

A novel feature of the Massachusetts data is the ability to link teachers to not only their educator preparation provider but also their specific program of study. We also find that individual programs in an institution vary considerably in the effectiveness of their teacher candidates. Providers and programs jointly explain about 10% of the variation in teacher effectiveness in ELA and 10–25% of the variation in math. In each case, the variability in average teacher effectiveness among programs within providers is similar or greater than the variability across providers. Using summative performance ratings under the Massachusetts educator evaluation framework, we found the same general pattern with less variability across providers and programs. Our findings on the variability of average educator effectiveness within institutions implies that the practices of specific programs are empirically important. Indeed, programs differ from others in their institution at least as much as institutions differ from one another. Practitioners and policymakers may wish to examine both program and institution data when evaluating educator preparation programs or making judgments about best practices. Louisiana and Tennessee, which have program evaluations that include a number of workforce outcomes, include some breakout by individual programs in their public reporting (Louisiana Board of Regents, 2014; Tennessee State Board of Education, 2015). However, the analysis of individual programs exacerbates the already limited sample sizes by splitting institutions into even smaller groups. Relying solely on program-level indicators may be infeasible for many smaller institutions and programs.

Despite the variability in educator effectiveness across providers, the exact rankings are sensitive to modeling decisions and we identify few programs that differ in a statistically significant way from the average of the provider indicators. Although our test of statistical significance using a 5% critical value is conservative, precisely ranking institutions is a common problem when evaluating educator preparation programs (Koedel et al., 2015b; Mihaly et al., 2013; von Hippel et al., 2016). The data set we use includes completers and student achievement data for five cohorts and up to 5 years. During this period, many institutions did not contribute large numbers

of teachers to our analysis; in fact, more than half of the providers in each of the value-added samples provided fewer than 50 completers. In addition to the general problems inherent in evaluating providers with few completers, there are a few additional challenges specific to the Massachusetts context. The state hosts a number of providers that disproportionately serve specific school systems or sectors. The inability to observe these teachers in a variety of schools with large numbers of teachers from different programs confounds attempts to separate the effectiveness of their completers from the effectiveness of the schools in which they serve. This is true for both value added, where research indicates that many of the specific charter schools networks staffed by small and specialized providers are more effective than nearby public schools (Abdulkadiroglu et al., 2011; Angrist et al., 2013), and for the summative performance data, where there is an important local role in the implementation of the evaluation system. The sensitivity of these results suggest some caution when interpreting indicators for providers with groups of completers that are highly concentrated in a small number of school districts or have other atypical school staffing patterns.

In our analysis, we also find variability in teacher retention across providers. Notably, the organizations that produced effective teachers were not always those whose teachers remained in the profession the longest. In **Table 16**, we show the correlations among the program indicators using the value-added, summative performance, and attrition data. The teacher effectiveness indicators are positively correlated, with two of the three relationships statistically significant. The teacher effectiveness measures are *also* positively correlated with the teacher attrition indicators; that is, the institutions that graduate more effective teachers also tend to have higher rates of attrition from the workforce.

Table 16. Correlation of Institution Indicators

	Math Value Added	ELA Value Added	Summative Performance Rating
Math Value Added			
ELA Value Added	0.55***		
Summative Performance Rating	0.26	0.39**	
Attrition	0.38**	0.43***	0.07

Note: Correlations for institution indicators. The correlations are estimated using all available data for each pair of indicators. For the value-added indicators, we used the baseline achievement models; for the summative performance ratings data, we used models with district fixed effects; for the attrition models, we used the baseline hazard estimates. All pairwise correlations are weighted by the sum of the number of teachers in each sample. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

The divergence between the attrition and effectiveness measures has a few implications for assessing the contribution of providers to the educator workforce. First, each of our indicators adjusts the effectiveness of teachers for their years of teaching experience. Given that the programs with more effective graduates also have higher attrition rates, the unconditional differences in average effectiveness—those that do not control for teaching experience—are likely to be smaller than those reported here. Our estimates therefore likely overstate the overall student achievement gains that would result from altering the composition of new hires in Massachusetts. Nonetheless, Kane et al. (2008) demonstrate that even modest differences in average teaching effectiveness can compensate for large differences in retention. Second, teacher

turnover may have indirect effects on student achievement that are not captured by the value-added or summative assessment measures we consider. Ronfeldt et al. (2013) found that teacher turnover had negative effects on student achievement in subsequent school years. Researchers have also observed a link between teacher turnover and school climate, which appears to affect student achievement generally and the professional development of new teachers (Kraft & Papay, 2014; Johnson et al., 2012). The available empirical evidence, however, suggests that the magnitude of these indirect effects is likely small relative to the direct student achievement impacts we estimate. Differences in retention rates are unlikely to substantially change how policymakers view individual programs and institutions.

State agencies and national accreditation groups are increasingly considering including performance measures, such as those considered in this report, in the accreditation and program approval process. The new regulations approved by the Massachusetts Board of Elementary and Secondary Education include measures of teacher performance. This policy is similar to guidance from the U.S. Department of Education and the new standards from the Council for the Accreditation of Educator Preparation, both of which incorporate several different indicators of program effectiveness to evaluate educator preparation programs (Council for the Accreditation of Educator Preparation, 2015). In order to provide a summary measure of our findings across a subset of these indicators, we classify institutions into quartiles based on each of the estimates in our data set. No single program is in either the highest or lowest quartile for all performance measures. Among the teacher effectiveness indicators, six institutions are in the highest quartile for two of the three measures, while nine institutions are in the bottom quartile for two of the three effectiveness indicators. Highlighting the discrepancies between the effectiveness and retention indicators, only two programs are in the highest quartile for three of the four indicators, and only three institutions are in the bottom quartile for at least three of the four indicators.

The use of multiple indicators provides policy makers with more reliable information with which to assess program effectiveness. As is the case with individual teachers, leveraging multiple measures of performance may improve the reliability of evaluation systems (Kane et al., 2013). For many programs, we observed only a small number of teachers. Although the effectiveness measures we estimate are educationally meaningful in magnitude, they are imprecise. Increasing the number of performance measures helps isolate the underlying signal about program effectiveness. Furthermore, because existing measures of teacher effectiveness are only proxies for underlying teaching skills, using a variety of indicators also broadens the range of skills an evaluation system might consider. Emerging research on teacher effectiveness suggests that teaching is multidimensional and that important teaching skills may not be captured either by test-based measures or by classroom observational tools (Gershensen, 2016; Harris & Sass, 2014; Jackson, 2016). The use of additional information about teacher effectiveness, such as licensure tests or manager evaluations, may improve our ability to predict the effectiveness of a programs' graduates.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135.

Abdulkadiroglu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and flexibility in public schools: evidence from Boston's charters and pilots. *The Quarterly Journal of Economics*, *126*(2), 699–748.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, *34*(2), 275–318.

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.

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal. Applied Economics*, *5*(4), 1–27.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (National Bureau of Economic Research Working Paper No. 20657). Cambridge, MA: National Bureau of Economic Research.

Ballou, D., & Podgursky, M. (1998). The case against teacher certification. *The Public Interest*, (132), 17–29.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Bastian, K. C., & Henry, G. T. (2015). Teachers without borders: Consequences of teacher labor force mobility. *Educational Evaluation and Policy Analysis*, *37*(2), 163–183.

Bastian, K. C., Patterson, T. M., & Pan, Y. (2015). *UNC teacher quality research: 2015 teacher preparation program effectiveness report*. Chapel Hill, NC: Education Policy Initiative at Carolina.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16–29.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school Administrators on teacher retention decisions. *American Educational Research Journal*, *48*(2), 303–333.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, *1*(2), 176–216.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). The draw of home: how teachers' preferences for proximity disadvantage urban schools. *Journal of Policy Analysis and Management*, *24*(1), 113–132.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2012). *Estimating the effect of leaders on public sector productivity: The case of school principals* (No. 17803). Cambridge, MA: National Bureau of Economic Research.

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2008). *National Board certification and teacher effectiveness: Evidence from a random assignment experiment* (NBER Working Paper No. 14022). Cambridge, MA: National Bureau of Economic Research.

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633–2679.

Chetty, R., Friedman, J. N., & Rockoff, J. (2016). Using lagged outcomes to evaluate bias in value-added models. *The American Economic Review*, *106*(5), 393–399.

Clark, M. A., Chiang, H. S., Silva, T., Mcconnell, S., Sonnenfeld, K., Erbe, A., . . . Puma, M. (2013). *The effectiveness of secondary math teachers from Teach For America and the teaching fellows programs* (No. NCEE 2013-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Clotfelter, C. T., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*(5–6), 1352–1370.

Clotfelter, C. T., Ladd, H., & Vigdor, J. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources*, *41*(4), 778–820.

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

Clotfelter, C., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *North Carolina Law Review, 85*, 1345–1380.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review, 40*, 76–92.

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification: final report* (No. NCEE 2009-4043). Washington, DC: National Center for Education Evaluation and Regional Assistance, National Center for Education Statistics, U.S. Department of Education.

Cook, J. B., & Mansfield, R. K. (2016). Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers. *Journal of Public Economics, 140*(8), 51–72.

Council for the Accreditation of Educator Preparation. (2015). *CAEP accreditation standards*. Washington, DC: Council for the Accreditation of Educator Preparation.

Cowan, J., & Goldhaber, D. (2016). National Board certification and teacher effectiveness: evidence from Washington state. *Journal of Research on Educational Effectiveness, 9*(3), 233–258.

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The effects of Teach for America on students: Findings from a national evaluation*. Princeton, NJ: Mathematica Policy Research.

Donaldson, M. L., & Johnson, S. M. (2011). Teach for America teachers: how long do they teach? Why do they leave? *Phi Delta Kappan, 93*(2), 47–51.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2013). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy, 1*(1), 19–27.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: should proportionality matter? *Educational Policy, 30*(3), 465–500.

Figlio, D. N., & Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics, 89*(2–3), 381–394.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*(4–5), 837–851.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., … Borman, G. D. (2016). *Focusing on mathematical knowledge: the impact of content-intensive teacher professional development* (No. NCEE 2016-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, *34*(2), 509-543.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125–149.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: the impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management*, *25*(1), 75–96.

Glazerman, S., & Seifullah, A. (2010). An evaluation of the Teacher Advancement Program (TAP) in Chicago: year two impact report. *Mathematica Policy Research*.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, 2(1), 50-55.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *The Review of Economics and Statistics*, *89*(1), 134–150.

Goldhaber, D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145.

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

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein falsification test": does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, *8*(1), 8–34.

Goldhaber, D., & Cowan, J. (2014). Excavating the teacher pipeline: Teacher preparation programs and teacher attrition. *Journal of Teacher Education*, *65*(5), 449–462.

Goldhaber, D., Cowan, J., & Walch, J. (2013a). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review*, *36*, 216–228.

Goldhaber, D., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *Journal of Policy Analysis and Management*, *30*(1), 57–87.

Goldhaber, D., Krieg, J. M., & Theobald, R. (2014). Knocking on the door to the teaching profession? Modeling the entry of prospective teachers into the workforce. *Economics of Education Review*, *43*, 106–124.

Goldhaber, D., Krieg, J., & Theobald, R. (2016). Does the match matter? Exploring whether student teaching experiences affect teacher effectiveness and career paths. CALDER Working Paper 149.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, *44*(5), 293–307.

Goldhaber, D., Liddle, S., & Theobald, R. (2013b). The gateway to the profession: assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29–44.

Goldhaber, D. and Walch, J. (2012). Strategic pay reform: a student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review*, 31(6), 1067-1083.

Goldhaber, D., Walch, J., & Gabele, B. (2013c). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, *1*(1), 28–39.

Goldring, R., Taie, S., & Riddles, M. (2014). *Teacher attrition and mobility: results from the 2012–13 Teacher Follow-Up Survey* (No. NCES 2014-077). Washington, DC: National Center for Education Statistics, U.S. Department Of Education.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: the relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445–470.

Guarino, C. M., Santibañez, L., & Daley, G. A. (2006). Teacher recruitment and retention: a review of the recent empirical literature. *Review of Educational Research*, *76*(2), 173–208.

Guarino, C.M., Reckase, M., Stacy, B., & Wooldridge, J. (2015). A comparison of student growth percentile and value-added models of teacher performance. *Statistics and Public Policy*, *2*(1), e1034820.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1), 117–156.

Hansen, M., Backes, B., & Brady, V. (2016). Teacher attrition and mobility during the Teach for America clustering strategy in Miami-Dade County Public Schools. *Educational Evaluation and Policy Analysis*, *38*(3), 495–516.

Hansen, M., & Sass, T. R. (2015). *Performance estimates of Teach for America teachers in Atlanta metropolitan area school districts* (No. 145). Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research.

Hanushek, E. (1986). The economics of schooling - production and efficiency in public-schools. *Journal of Economic Literature* 24: 1141-1177.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 267-271.

Harris, D. N., & Sass, T. R. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management: [the Journal of the Association for Public Policy Analysis and Management]*, *28*(1), 55–80.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7-8), 798–812.

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

Henry, G. T., Purtell, K. M., Bastian, K. C., Fortner, C. K., Thompson, C. L., Campbell, S. L., . . . Patterson, K. M. (2014). The effects of teacher entry portals on student achievement. *Journal of Teacher Education*, *65*(1), 7–23.

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education*, 330-351.

Imazeki, J. (2005). Teacher salaries and teacher attrition. *Economics of Education Review*, *24*(4), 431–449.

Jackson, C. K. (2012). School competition and teacher labor markets: evidence from charter school entry in North Carolina. *Journal of Public Economics*, *96*(5-6), 431–448.

Jackson, C. K. (2014). Teacher quality at the high school level: the importance of accounting for tracks. *Journal of Labor Economics*, *32*(4), 645–684.

Jackson, C. K. (2016). *What do test scores miss? The importance of teacher effects on non-test score outcomes* (No. 22226). Cambridge, MA: National Bureau of Economic Research.

Jacob, B. A. (2011). Do principals fire the worst teachers? *Educational Evaluation and Policy Analysis*, *33*(4), 403–434.

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement: quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, 39(1), 50-79.

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

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: the effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*(10), 1–39.

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

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *The Journal of Human Resources*, *46*(3), 587–613.

Kinsler, J. (2012). Beyond levels and growth: Estimating teacher value-added and its persistence. *The Journal of Human Resources*, *47*(3), 722–753.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, *6*(1), 18–42.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015a). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

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

Kraft, M. A., & Gilmour, A. F. (2016). Revisiting *The Widget Effect*: Teacher evaluation reforms and the distribution of teacher effectiveness. Brown University Working Paper.

Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, *36*(4), 476–500.

Krieg, J. M., Theobald, R., & Goldhaber, D. (2016). A foot in the door: Exploring the role of student teaching assignments in teachers' initial job placements. *Educational Evaluation and Policy Analysis*, *38*(2), 364–388.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37–62.

Larsen, B. (2015). Occupational licensing and quality: Distributional and heterogeneous effects in the teaching profession. Unpublished manuscript.

Lincove, J. A., Osborne, C., Dillon, A., & Mills, N. (2014). The politics and statistics of value-added modeling for accountability of teacher preparation programs. *Journal of Teacher Education*, *65*(1), 24–38.

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*, *36*(4), 457–475.

Louisiana Board of Regents. (2014). *Louisiana teacher preparation program factbook*. Baton Rouge, LA: Louisiana Board of Regents. Retrieved from http://www.uno.edu/coehd/ncate-2015/Exhibits/Q79354-1.4.k4-Final-Teacher-Prep-Fact-Book.pdf

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics, 33*(3), 751–788.

Massachusetts Department of Elementary and Secondary Education. (2015a). *Guidelines for Professional Standards for Teachers*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2015b). *Massachusetts educator evaluation. Rating educator performance: The summative performance rating*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2015c). *Massachusetts model system for educator evaluation, part I: District-level planning and implementation guide*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2015d). *Massachusetts model system for educator evaluation, part II: School-level planning and implementation guide*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2015e). *Massachusetts model system for educator evaluation, part III: Guide to rubrics and model rubrics for superintendent, administrator, and teacher*. Malden, MA: Author.

Mihaly, K., McCaffrey, D., Sass, T. R., & Lockwood, J. R. (2013). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy*, *8*(4), 459–493.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. New York, NY: Author.

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal. Applied Economics*, *6*(2), 127–151.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163–193.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119.

Papay, J. P., West, M. R., Fullerton, J. B., & Kane, T. J. (2012). Does an urban teacher residency increase student achievement? Early evidence from Boston. *Educational Evaluation and Policy Analysis*, *34*(4), 413–434.

Podgursky, M., Monroe, R., & Watson, D. (2004). The academic quality of public school teachers: an analysis of entry and exit behavior. *Economics of Education Review*, *23*(5), 507–518.

Protik, A., Walsh, E., Resch, A., Isenberg, E., & Kopa, E. (2013). Does tracking of students bias value-added estimates for teachers? In *Association of Education Finance and Policy Conference*.

Provasnik, S., & Dorfman, S. (2005). *Mobility in the teacher workforce* (No. NCES 2005–114). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Redding, C., & Smith, T. M. (2016). Easy in, easy out: Are alternatively certified teachers turning over at increased rates? *American Educational Research Journal*, *53*(4), 1086–1125.

Reininger, M. (2012). Hometown disadvantage? It depends on where you're from: Teachers' location preferences and the implications for staffing schools. *Educational Evaluation and Policy Analysis*, *34*(2), 127–145.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247–252.

Rockoff, J. E. (2008). *Does mentoring reduce turnover and improve skills of new employees? Evidence from teachers in New York City*. National Bureau of Economic Research.

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

Ronfeldt, M., & Campbell, S. L. (2016). Evaluating teacher preparation using graduates' observational ratings. *Educational Evaluation and Policy Analysis*. Advance online publication.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, *50*(1), 4–36.

Ronfeldt, M., Schwartz, N., & Jacob, B. A. (2014). Does pre-service preparation matter? Examining an old question in new ways. *Teachers College Record*, *116*(10), 1–46.

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

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

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

Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal*, *41*(3), 681–714.

Springer, M. G., Ballou, D., Hamilton, L., Le, V. N., Lockwood, J. R., McCaffrey, D. F., ... & Stecher, B. M. (2011). Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching (POINT). *Society for Research on Educational Effectiveness*.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis*, *38*(2), 199–221.

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

Tennessee State Board of Education. (2015). *Tennessee teacher preparation report card: State profile*. Nashville, TN: Tennessee State Board of Education.

U.S. Department of Education. (2009). *Growth models: non-regulatory guidance*. Washington, D.C.: U.S. Department of Education.

U.S. Department of Education. (2013). *Preparing and credentialing the nation's teachers: The Secretary's ninth report on teacher quality*. Washington, DC: Office of Postsecondary Education, U.S. Department of Education.

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

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect*. New York, NY: TNTP.

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

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, *100*, 61–78.

Wooldridge, J. M. (2010). *Correlated random effects models with unbalanced panels*. Unpublished manuscript.

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach for America in high school. *Journal of Policy Analysis and Management*, *30*(3), 447–469.

Appendix A. Additional Methodological Details and Robustness Checks

Additional Descriptive Statistics by Institution

In **Table A.1**, we provide counts of the number of teachers by institution and analysis sample. In order to protect the anonymity of the institutions, we have masked the exact number of completers. Instead, we group institutions into groups of <15, 15-50, and 50+.

Table A.1. Sample Sizes by Preparation Institution and Analytic Sample

Organization	Attrition	Summative Evaluation	VAM (Math)	VAM (ELA)
Program 1	50+	50+	15-50	15-50
Program 2	<15	<15	<15	<15
Program 3	15-50	50+	15-50	15-50
Program 4	50+	50+	50+	50+
Program 5	50+	50+	15-50	15-50
Program 6	50+	50+	50+	50+
Program 7	50+	50+	15-50	15-50
Program 8	50+	50+	50+	50+
Program 9	<15	<15	<15	<15
Program 10	50+	50+	15-50	15-50
Program 11	<15	15-50	<15	<15
Program 12	<15	<15	<15	<15
Program 13	50+	50+	50+	50+
Program 14	15-50	15-50	<15	<15
Program 15	<15	<15	<15	<15
Program 16	15-50	15-50	<15	<15
Program 17	15-50	<15	<15	<15
Program 18	50+	50+	15-50	15-50
Program 19	<15	<15	<15	<15
Program 20	<15	<15	<15	<15
Program 21	15-50	15-50	15-50	15-50
Program 22	50+	50+	50+	50+
Program 23	50+	50+	15-50	15-50
Program 24	50+	50+	50+	50+
Program 25	50+	50+	50+	15-50
Program 26	15-50	50+	<15	<15
Program 27	<15	<15	<15	<15
Program 28	50+	50+	50+	50+

Organization	Attrition	Summative Evaluation	VAM (Math)	VAM (ELA)
Program 29	50+	50+	50+	50+
Program 30	50+	50+	50+	50+
Program 31	15-50	15-50	<15	<15
Program 32	50+	50+	15-50	<15
Program 33	50+	50+	15-50	15-50
Program 34	50+	50+	15-50	15-50
Program 35	<15	<15	<15	<15
Program 36	<15	<15	<15	<15
Program 37	50+	50+	15-50	15-50
Program 38	<15	<15	<15	<15
Program 39	50+	50+	50+	15-50
Program 40	<15	<15	<15	<15
Program 41	50+	50+	<15	15-50
Program 42	<15	<15	<15	<15
Program 43	15-50	15-50	<15	<15
Program 44	50+	50+	15-50	15-50
Program 45	<15	<15	<15	<15
Program 46	50+	50+	15-50	15-50
Program 47	<15	<15	<15	<15
Program 48	15-50	15-50	<15	<15
Program 49	50+	50+	50+	50+
Program 50	15-50	15-50	<15	<15
Program 51	50+	50+	15-50	15-50
Program 52	<15	<15	<15	<15
Program 53	<15	<15	<15	<15
Program 54	50+	50+	15-50	15-50
Program 55	50+	50+	50+	50+
Program 56	50+	50+	50+	50+
Program 57	50+	50+	50+	50+
Program 58	15-50	15-50	<15	<15
Program 59	50+	50+	50+	50+
Program 60	15-50	15-50	<15	15-50
Program 61	15-50	15-50	<15	<15
Program 62	50+	50+	15-50	15-50
Program 63	50+	50+	15-50	15-50
Program 64	<15	<15	<15	<15
Program 65	50+	50+	15-50	15-50
Program 66	15-50	15-50	<15	<15

Organization	Attrition	Summative Evaluation	VAM (Math)	VAM (ELA)
Program 67	50+	50+	15-50	15-50
Program 68	50+	50+	50+	50+

Note: Table contains number of completers included in each of the analysis samples. For the evaluation and value-added analyses, we use completers from the 2010 to 2014 cohorts. For the attrition analysis, we do not have data on the status of teachers in the 2015–2016 school year. We therefore rely on data from 2011 to 2015 and only include teachers in the 2010–2013 cohorts. The summative performance analysis includes the 2010–2014 completer cohorts and uses data from the 2014–2015 school years. Programs with fewer than 15 completers in the given sample are not included in the analysis. Exact counts are masked to preserve anonymity. Abbreviations: VAM = value-added model. Names of EPPs have been removed from report to protect the confidentiality of programs in Massachusetts.

We present average school demographic information by preparation institution for teachers in the program completer sample in **Table A.2**. As with the summary statistics by licensure pathway, we use average school aggregates for teachers in the attrition sample. The table indicates that there is substantial heterogeneity across organizations in the characteristics of schools in which their graduates teach. For instance, the average subsidized lunch rate is 47%, but graduates of three programs (21, 51, and 67) teach in schools with mean subsidized lunch rates exceeding 70%. We see similar variation in the proportion of students who are English language learners; for example, graduates of Programs 51, 54, and 67 teach in schools with average English language learner rates over 20%. On the other hand, more than half of all institutions have graduates who work in schools where fewer than 10% of students are English language learners. The summary statistics in Table 4 indicate that there are meaningful differences across institutions in the characteristics of schools in which their graduates teach. These patterns appear in other states as well, partially because some alternative preparation providers are specifically designed to place graduates in certain types of teaching positions, and also because teachers in traditional preparation programs tend to attend school, complete their student teaching, and obtain teaching positions near where they grew up (Krieg et al., 2016; Loeb et al., 2005; Reininger, 2012).

Table A.2. Average School Demographics by Preparation Institution

Organization	Percent Hispanic	Percent Black	Percent Asian	Percent Free- or Reduced-Price Lunch	Percent English Language Learner	Percent Special Education
Program 1	23.7	8.8	4.1	48.9	8.4	17.8
Program 3	24.7	10.1	5.4	49.5	13.3	17.4
Program 4	25.0	18.4	8.1	52.1	14.6	19.1
Program 5	25.3	5.6	5.1	47.5	10.7	17.1
Program 6	22.2	22.8	8.2	54.5	15.8	19.0
Program 7	16.4	7.5	6.6	39.6	9.0	17.0
Program 8	14.6	5.9	6.8	32.8	8.0	17.0
Program 10	20.8	4.6	5.7	37.2	8.4	18.0
Program 11	26.3	13.0	4.1	47.2	10.8	20.0
Program 13	17.7	11.3	8.0	38.2	9.5	17.0
Program 14	20.0	10.3	6.0	48.1	14.1	16.3

Organization	Percent Hispanic	Percent Black	Percent Asian	Percent Free- or Reduced-Price Lunch	Percent English Language Learner	Percent Special Education
Program 16	6.9	4.2	8.0	17.6	4.3	16.2
Program 17	11.8	6.8	2.7	33.2	6.3	16.0
Program 18	16.4	11.6	4.4	41.3	8.9	18.6
Program 20	15.8	3.5	5.3	37.2	3.7	19.8
Program 21	44.5	41.9	2.5	85.5	16.4	20.8
Program 22	22.4	6.6	5.6	45.0	9.7	18.2
Program 23	10.1	13.4	4.5	38.9	7.2	16.5
Program 24	9.4	9.2	4.0	38.6	5.4	17.0
Program 25	26.9	12.9	7.7	52.2	14.2	17.9
Program 26	10.6	21.3	2.5	50.6	7.3	18.7
Program 28	25.8	12.5	7.8	47.0	11.7	17.8
Program 29	19.2	6.8	4.6	46.2	11.9	17.2
Program 30	19.8	14.4	8.2	42.6	12.2	18.7
Program 31	23.6	14.3	4.8	52.0	10.5	18.6
Program 32	9.5	12.8	13.2	43.8	11.0	16.6
Program 33	20.5	4.0	11.3	42.8	9.1	18.6
Program 34	24.0	17.1	7.2	49.3	17.0	16.9
Program 37	29.4	6.2	4.5	57.1	10.9	21.5
Program 38	25.3	11.4	2.2	50.3	5.7	16.4
Program 39	26.8	7.9	4.8	52.6	9.7	17.4
Program 41	26.2	24.7	9.1	52.1	14.0	18.2
Program 43	13.3	12.2	11.9	36.0	7.6	16.9
Program 44	12.0	13.0	4.4	41.4	7.2	15.5
Program 46	38.3	6.9	2.1	66.1	10.2	18.4
Program 48	15.4	14.6	6.3	36.4	8.5	16.7
Program 49	22.4	7.5	4.4	48.3	7.8	18.3
Program 50	17.6	11.2	3.6	42.4	7.9	16.3
Program 51	56.4	21.0	3.5	82.6	21.6	17.7
Program 54	35.5	13.0	8.2	62.6	26.9	16.7
Program 55	15.2	13.8	2.3	58.5	6.6	17.6
Program 56	17.0	9.1	8.3	37.2	9.9	16.9
Program 57	25.2	8.2	4.8	51.9	10.5	18.9
Program 58	5.3	4.0	6.3	14.2	2.7	16.7
Program 59	20.5	5.6	5.6	46.1	9.0	18.6
Program 60	24.5	6.9	3.3	52.5	7.6	18.6
Program 61	29.7	8.9	4.2	58.6	12.7	17.3

Organization	Percent Hispanic	Percent Black	Percent Asian	Percent Free- or Reduced-Price Lunch	Percent English Language Learner	Percent Special Education
Program 62	19.8	15.7	8.4	42.7	9.0	16.4
Program 63	8.9	9.9	4.3	36.1	6.5	17.1
Program 65	9.1	5.7	1.5	49.3	3.5	17.8
Program 66	21.7	9.3	7.8	42.4	7.8	18.2
Program 67	38.2	38.6	6.7	77.3	28.1	20.5
Program 68	30.9	7.9	2.9	59.6	10.5	18.2

Note: Table presents average school characteristics by institution for teachers in attrition sample using data from 2011 to 2014. Observations are at the teacher-year level.

Value-Added Models

The primary value-added research design we employ can be summarized by the following regression equation:

$$Y_{ijst} = X_{ijst}\beta + P_{jt}\delta + \alpha_s + \epsilon_{ijst}. \quad \text{(A.1)}$$

In Eq. (A.1), the subscript *i* indexes students, *j* indexes teachers, *s* indexes school, and *t* indexes year. We regress student achievement on the MCAS or PARCC for each student on a set of variables *X* that includes prior achievement, academic and demographic characteristics, and peer and school characteristics. The teacher characteristics *P* include descriptions of their preparation background (e.g., licensure pathway, preparation institution, program). We are interested in the teacher preparation variables *P* and their associated standard errors.[37] In addition to these variables, we include school fixed effects α_s in some specifications, which ensure that comparisons of teachers from different programs are made only within schools. Some specifications also include controls for students' class assignments to ensure that teachers are compared only to others teaching students in the same track. Similar specifications of the student achievement equation provide the foundation for much of the research literature on teacher credentials and preparation (Boyd et al., 2009; Goldhaber et al., 2013; Mihaly et al., 2013). We present the provider estimates using the full range of model specifications in **Table A.3**.

Estimation of the value-added model Eq. (A.1), like the other empirical models we estimate for teacher evaluation results and retention, produces both a point estimate of a particular program's average teacher effectiveness and a standard error that describes the uncertainty of the estimate. This uncertainty results from the limitations of the data. We observe only a limited number of teachers in a limited number of years per program. In each year, as few as 20 students may inform the estimate of an individual teacher's effectiveness. Prior research has documented that teacher performance varies from year to year due to the small number of student observations available per teacher, changes in classroom chemistry, and changes in the effectiveness of individual teachers over time (Chetty et al., 2014a; Goldhaber & Hansen, 2013; McCaffrey et al.,

[37] In each of our analyses, we estimate the program effects using the parameterization described by Mihaly et al. (2010). The coefficients are interpretable as the deviation from the mean program in Massachusetts.

2009). Typically, the resulting uncertainty in estimates of program effectiveness is quantified by the use of confidence intervals, which describe a range that would be expected to cover a certain percentage of replications of the estimation procedure from different samples. In the social sciences, researchers typically use a 95% confidence interval, which is expected to cover 95% of the estimates observed from repeated trials. Some accountability systems for educator preparation programs rely on the use of confidence intervals and identify programs whose confidence intervals exclude certain critical values, such as the mean of the program effects (Lincove et al., 2014). This approach is similar to hypothesis testing in social science research. However, the choice of a particular confidence interval to identify exemplary or ineffective programs is also implicitly a choice for the sensitivity of the test. Selecting a higher level for the confidence interval reduces the likelihood of inadvertently identifying a program as ineffective, but it necessarily increases the probability of failing to detect programs that, in fact, deviate from the mean.

Table A.3. Alternative Specifications for Value Added Models

	(1)	(2)	(3)	(4)	(5)	(6)
	Math			ELA		
Program 1	-0.003	0.120***	0.077**	-0.047**	-0.024	-0.006
	(0.030)	(0.031)	(0.036)	(0.022)	(0.034)	(0.036)
Program 3	-0.060	-0.076*	-0.124**	-0.072*	-0.147***	-0.178***
	(0.039)	(0.044)	(0.049)	(0.038)	(0.042)	(0.041)
Program 4	-0.012	0.040*	0.032	-0.044*	-0.001	0.026
	(0.031)	(0.024)	(0.022)	(0.025)	(0.025)	(0.024)
Program 5	-0.058	-0.028	-0.008	-0.057**	-0.012	0.021
	(0.057)	(0.046)	(0.036)	(0.027)	(0.035)	(0.034)
Program 6	-0.006	0.010	-0.005	-0.008	-0.027	0.021
	(0.021)	(0.021)	(0.023)	(0.022)	(0.021)	(0.022)
Program 7	-0.025	-0.022	-0.054**	-0.021	0.024	-0.015
	(0.034)	(0.028)	(0.025)	(0.026)	(0.036)	(0.054)
Program 8	-0.045***	-0.036**	0.004	-0.021	-0.025	0.008
	(0.015)	(0.018)	(0.018)	(0.021)	(0.021)	(0.022)
Program 10	-0.020	-0.045	0.007	-0.043	0.058*	-0.010
	(0.044)	(0.050)	(0.066)	(0.041)	(0.031)	(0.041)
Program 13	0.017	-0.029	-0.036	0.036*	0.032*	0.042**
	(0.025)	(0.021)	(0.025)	(0.021)	(0.019)	(0.019)
Program 18	0.037	0.082	0.071	-0.068	0.048	0.043
	(0.050)	(0.050)	(0.055)	(0.045)	(0.043)	(0.043)
Program 21	0.222***	-0.123*	-0.023	0.129***	-0.185**	-0.109
	(0.040)	(0.063)	(0.063)	(0.044)	(0.077)	(0.092)
Program 22	-0.064***	-0.052***	-0.026	0.005	-0.013	-0.015
	(0.018)	(0.019)	(0.021)	(0.018)	(0.017)	(0.019)

	(1)	(2)	(3)	(4)	(5)	(6)
	Math			ELA		
Program 23	-0.005	0.047**	0.046**	0.006	-0.031	0.033
	(0.028)	(0.020)	(0.020)	(0.029)	(0.031)	(0.031)
Program 24	-0.017	0.016	0.035**	-0.021	-0.004	0.021
	(0.013)	(0.015)	(0.015)	(0.014)	(0.015)	(0.016)
Program 25	0.025	0.014	0.026	-0.013	-0.029	-0.004
	(0.029)	(0.022)	(0.020)	(0.033)	(0.034)	(0.042)
Program 28	0.032	0.043**	0.035*	0.008	0.023	0.008
	(0.026)	(0.018)	(0.018)	(0.037)	(0.028)	(0.028)
Program 29	-0.011	-0.002	-0.066**	0.007	-0.032	0.010
	(0.023)	(0.027)	(0.032)	(0.020)	(0.023)	(0.024)
Program 30	0.048*	0.082***	0.054*	0.001	-0.037	0.017
	(0.027)	(0.029)	(0.029)	(0.025)	(0.025)	(0.024)
Program 32	-0.109***	0.006	-0.056			
	(0.042)	(0.045)	(0.037)			
Program 33	0.064***	0.061*	0.003	0.002	-0.002	0.026
	(0.024)	(0.036)	(0.033)	(0.028)	(0.026)	(0.025)
Program 34	-0.070	-0.064	0.079	0.021	0.001	0.088
	(0.057)	(0.056)	(0.089)	(0.055)	(0.074)	(0.062)
Program 37	-0.027	0.109*	0.019	0.075	0.104	-0.181**
	(0.042)	(0.058)	(0.065)	(0.057)	(0.086)	(0.087)
Program 39	-0.023	-0.056	-0.102*	-0.012	-0.058	-0.096**
	(0.029)	(0.035)	(0.053)	(0.034)	(0.036)	(0.047)
Program 41				0.007	-0.043	0.014
				(0.024)	(0.030)	(0.030)
Program 44	0.025	0.040*	0.037	-0.021	0.004	0.004
	(0.032)	(0.022)	(0.028)	(0.038)	(0.025)	(0.025)
Program 46	-0.012	-0.019	-0.043	0.009	0.055	0.063*
	(0.040)	(0.048)	(0.045)	(0.033)	(0.037)	(0.037)
Program 49	-0.025	0.002	0.001	-0.044*	-0.013	0.019
	(0.020)	(0.023)	(0.019)	(0.024)	(0.022)	(0.021)
Program 51	0.185**	0.074	-0.015	0.298***	0.159*	-0.012
	(0.073)	(0.061)	(0.046)	(0.079)	(0.083)	(0.048)
Program 54	-0.030	0.011	-0.033	0.069**	0.088*	0.085**
	(0.050)	(0.058)	(0.049)	(0.034)	(0.053)	(0.036)
Program 55	-0.001	-0.001	0.018	-0.019	0.054**	0.062**
	(0.025)	(0.022)	(0.026)	(0.022)	(0.024)	(0.030)
Program 56	0.012	-0.019	0.011	0.021	0.045**	0.062***

	(1)	(2)	(3)	(4)	(5)	(6)
	Math			ELA		
	(0.019)	(0.021)	(0.024)	(0.017)	(0.018)	(0.019)
Program 57	-0.059***	-0.026	0.030	-0.042**	0.010	0.014
	(0.022)	(0.026)	(0.027)	(0.021)	(0.026)	(0.029)
Program 59	-0.008	-0.054**	-0.010	-0.012	0.014	0.026
	(0.028)	(0.024)	(0.024)	(0.015)	(0.022)	(0.020)
Program 60				0.028	0.071**	0.066**
				(0.024)	(0.034)	(0.030)
Program 62	-0.034	0.001	-0.013	0.018	-0.045	-0.102**
	(0.044)	(0.056)	(0.029)	(0.053)	(0.050)	(0.051)
Program 63	0.086***	0.043	0.096**	0.090**	0.134***	0.069
	(0.033)	(0.037)	(0.045)	(0.038)	(0.047)	(0.051)
Program 65	0.028	-0.148**	-0.036	-0.155***	-0.094**	-0.075*
	(0.050)	(0.070)	(0.071)	(0.034)	(0.038)	(0.040)
Program 67	-0.031	0.022	0.002	-0.102**	-0.105*	-0.055
	(0.039)	(0.034)	(0.034)	(0.051)	(0.058)	(0.050)
Program 68	-0.025	-0.024	-0.030	-0.008	0.003	0.007
	(0.024)	(0.023)	(0.029)	(0.025)	(0.024)	(0.023)
School FE	N	Y	N	N	Y	N
Track FE	N	N	Y	N	N	Y
Model 1 Correlation	1.000	0.104	0.186	0.619	0.057	-0.043
Model 2 Correlation	0.104	1.000	0.537	0.234	0.493	0.181
Model 3 Correlation	0.186	0.537	1.000	0.034	0.345	0.472
Model 4 Correlation	0.619	0.234	0.034	1.000	0.473	0.059
Model 5 Correlation	0.057	0.493	0.345	0.473	1.000	0.493
Model 6 Correlation	-0.043	0.181	0.472	0.059	0.493	1.000

Note: Table presents alternative specifications for value added models. All models include controls for potential teaching experience, a cubic polynomial in prior achievement in math and ELA interacted with grade, gender, race/ethnicity, subsidized lunch status, and the classroom and school aggregates of these variables. Track fixed effects indicate school-grade-level cells. Correlations in the table indicate the correlation of estimated institution effects with the results from the specified model. Standard errors are clustered by teacher. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Summative Performance Ratings

To estimate institution indicators using the summative performance ratings, we estimate models that control for teacher experience and school characteristics. The inclusion of school characteristics is motivated by research suggesting that classroom composition influences observational evaluations, which are one component of the Massachusetts educator evaluation framework (Steinberg and Garret, 2016; Whitehurst et al., 2014). We estimate the regression model

$$R_{jst} = X_{jst}\gamma + T_{jt}\beta + P_j\delta + \alpha_s + \epsilon_{jt}. \quad \text{(A.2)}$$

where R is the summative rating measured on a 1 – 4 scale, X_{jst} is a vector of school characteristics, T_{jt} is a vector of experience variables, and P_j indicates a teacher's program or pathway. We additionally estimate models that includes school or district fixed effects. This approach follows that used by Ronfeldt and Campbell (2016) in their analysis of preparation programs in Tennessee. In **Table A.4**, we present coefficients from the full set of models.

Bastian et al. (2015) estimate a model similar to Eq. (A.2) by ordered logistic regression that includes controls for experience and school demographics. The ordered logistic model relies on similar assumptions as the linear regression, but better accounts for the fact that the summative performance ratings are not measured on a natural scale. These models allow for the possibility that a movement from unsatisfactory to needs improvement may not connote the same difference in teacher quality as a movement from proficient to exemplary. As we demonstrate below, the most important modeling decisions involve adjustments for school or district heterogeneity rather than the specific distributional assumptions. Results for similar specifications are highly correlated regardless of whether we estimate them using linear regression or models for ordered data. This finding motivates our decision to present estimates from linear models in the main text.

We pursue two extensions of the baseline linear model. First, the ratings data are properly ordinal rather than cardinal: they enumerate performance categories, but there is not necessarily any meaning to the distance between performance categories. For example, there are more teachers classified as needs improvement in our sample than are classified as exemplary. The difference in average effectiveness between the needs improvement and proficient ratings is likely smaller than the difference between proficient and exemplary. We can account for this kind of discrete data using an ordinal probit model, which is similar to the approach taken by Bastian et al. (2015) in an analysis of educator preparation programs in North Carolina. The ordered probit models treat the ratings as ordinal data: there is a natural ranking of the performance categories, but they do not provide information on "how much" better a teacher in the needs improvement category performs relative to one in the proficient category. Instead, the ordered probit model is premised on the idea that there is an underlying scale of teacher effectiveness. We assume that raters have in mind a series of cut points that determine which rating a particular candidate receives.

The second specification concerns the degree to which districts distinguish between teachers of varying effectiveness. The basic ordered probit model with controls for district staffing and school characteristics allow for rating standards to shift across districts, but they assume that the

relative distances between rating categories are constant. However, a careful inspection of the ratings distributions in Figure 2 indicates that districts do not just rate teachers higher or lower. Rather, it appears that some districts assign nearly every teacher proficient whereas others assign more teachers to both the needs improvement and exemplary ratings. We therefore additionally estimate models that allow for district characteristics to affect both average scores and the cut scores that determine which rating a teacher receives. We implement this using the generalized ordered probit model described by Pudney and Shields (2000). The generalized ordered probit models the thresholds as a linear function of the district characteristics (including the staff composition). This specification allows districts to additionally vary in their sensitivity to differences in teacher quality – that is the extent to which they use the rating scale to discriminate among teachers with varying effectiveness. As the results in Table A.2 indicate, results from this model are very similar to those that simply use district fixed effects. We conclude that the exact model specification is less important in this context than the decision to account for variation in ratings across districts and schools.

Table A.4. Alternative Specifications of Summative Performance Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Program 1	0.015	0.012	-0.003	0.017	-0.002	0.071	0.084
	(0.018)	(0.019)	(0.021)	(0.022)	(0.026)	(0.092)	(0.108)
Program 3	-0.010	0.002	-0.007	-0.027	-0.042	-0.088	-0.084
	(0.036)	(0.036)	(0.039)	(0.049)	(0.053)	(0.175)	(0.151)
Program 4	-0.005	-0.013	-0.001	-0.008	0.001	-0.027	-0.036
	(0.018)	(0.017)	(0.019)	(0.024)	(0.024)	(0.092)	(0.090)
Program 5	-0.014	0.020	0.013	0.001	0.011	0.024	0.013
	(0.022)	(0.021)	(0.023)	(0.033)	(0.035)	(0.118)	(0.102)
Program 6	-0.002	-0.026*	-0.041***	-0.019	-0.040**	-0.073	-0.060
	(0.014)	(0.014)	(0.015)	(0.018)	(0.020)	(0.070)	(0.067)
Program 7	-0.033*	-0.038*	-0.028	-0.023	-0.008	-0.094	-0.101
	(0.020)	(0.020)	(0.021)	(0.028)	(0.029)	(0.110)	(0.123)
Program 8	-0.037**	-0.033**	-0.026	-0.038*	-0.032	-0.149*	-0.151**
	(0.015)	(0.015)	(0.017)	(0.020)	(0.022)	(0.077)	(0.074)
Program 10	-0.049**	-0.033	-0.030	-0.076**	-0.077**	-0.269**	-0.281**
	(0.025)	(0.025)	(0.024)	(0.033)	(0.031)	(0.123)	(0.122)
Program 11	0.044	0.049	0.059	-0.078	-0.074	-0.292	-0.235
	(0.060)	(0.058)	(0.061)	(0.079)	(0.077)	(0.282)	(0.282)
Program 13	-0.024*	-0.022	-0.026*	-0.008	-0.015	-0.036	-0.044
	(0.014)	(0.014)	(0.014)	(0.018)	(0.018)	(0.070)	(0.067)
Program 14	-0.007	-0.019	-0.023	-0.024	-0.021	-0.055	-0.028
	(0.049)	(0.052)	(0.049)	(0.052)	(0.050)	(0.212)	(0.201)
Program 16	0.018	0.039	0.018	0.064	0.067	0.237	0.287
	(0.069)	(0.064)	(0.052)	(0.078)	(0.074)	(0.318)	(0.252)
Program 18	-0.020	-0.003	-0.009	0.013	0.008	0.056	0.060

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(0.022)	(0.022)	(0.022)	(0.027)	(0.029)	(0.110)	(0.117)
Program 21	0.050	0.014	0.079	0.089	0.125	0.275	0.223
	(0.083)	(0.087)	(0.110)	(0.118)	(0.153)	(0.434)	(0.230)
Program 22	-0.015	0.004	0.001	0.009	0.005	0.034	0.038
	(0.011)	(0.012)	(0.013)	(0.017)	(0.018)	(0.067)	(0.062)
Program 23	0.094***	0.093***	0.097***	0.101***	0.104***	0.387***	0.421***
	(0.029)	(0.029)	(0.029)	(0.036)	(0.036)	(0.145)	(0.134)
Program 24	-0.047***	-0.018	-0.013	-0.027*	-0.027	-0.093	-0.090
	(0.010)	(0.011)	(0.012)	(0.015)	(0.017)	(0.057)	(0.057)
Program 25	0.008	0.008	0.011	-0.011	-0.017	-0.027	-0.029
	(0.026)	(0.025)	(0.025)	(0.033)	(0.034)	(0.117)	(0.097)
Program 26	-0.072*	-0.031	-0.007	-0.001	0.014	-0.026	-0.018
	(0.039)	(0.045)	(0.047)	(0.044)	(0.050)	(0.175)	(0.194)
Program 28	-0.014	0.010	0.006	0.010	0.005	0.037	0.030
	(0.014)	(0.013)	(0.015)	(0.020)	(0.022)	(0.077)	(0.078)
Program 29	0.024	0.000	0.014	0.006	0.018	0.040	0.059
	(0.017)	(0.017)	(0.017)	(0.021)	(0.022)	(0.087)	(0.086)
Program 30	0.028	0.029	0.011	0.033	0.011	0.137	0.171*
	(0.022)	(0.020)	(0.020)	(0.027)	(0.027)	(0.104)	(0.098)
Program 31	-0.033	-0.022	-0.085**	0.005	-0.069	-0.020	-0.043
	(0.037)	(0.036)	(0.043)	(0.030)	(0.045)	(0.160)	(0.319)
Program 32	-0.033	-0.022	-0.013	-0.028	-0.006	-0.103	-0.128
	(0.025)	(0.027)	(0.025)	(0.030)	(0.031)	(0.120)	(0.192)
Program 33	0.003	-0.005	0.003	-0.026	-0.008	-0.090	-0.080
	(0.023)	(0.025)	(0.024)	(0.033)	(0.034)	(0.124)	(0.119)
Program 34	-0.056**	-0.064**	-0.101***	-0.062**	-0.081**	-0.257**	-0.228**
	(0.025)	(0.026)	(0.028)	(0.031)	(0.035)	(0.114)	(0.112)
Program 37	0.059**	0.043	0.039	0.059*	0.059*	0.242**	0.235**
	(0.024)	(0.027)	(0.028)	(0.031)	(0.034)	(0.118)	(0.119)
Program 39	-0.004	-0.011	-0.018	-0.001	-0.003	-0.009	-0.006
	(0.022)	(0.023)	(0.025)	(0.027)	(0.030)	(0.112)	(0.104)
Program 41	-0.051*	-0.087***	-0.064*	-0.059**	-0.039	-0.238**	-0.241
	(0.030)	(0.028)	(0.034)	(0.029)	(0.040)	(0.120)	(0.173)
Program 43	0.024	-0.002	-0.005	-0.060	-0.097	-0.222	-0.247
	(0.052)	(0.051)	(0.059)	(0.072)	(0.075)	(0.284)	(0.269)
Program 44	-0.017	0.007	0.005	0.004	0.002	0.034	0.065
	(0.018)	(0.018)	(0.020)	(0.025)	(0.029)	(0.109)	(0.120)
Program 46	-0.017	-0.034	-0.052	-0.034	-0.032	-0.132	-0.108

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(0.034)	(0.035)	(0.035)	(0.041)	(0.043)	(0.149)	(0.131)
Program 48	0.013	-0.023	-0.003	-0.019	0.028	-0.075	-0.106
	(0.057)	(0.051)	(0.050)	(0.068)	(0.068)	(0.262)	(0.232)
Program 49	-0.021	-0.025*	-0.043***	-0.034*	-0.054***	-0.147**	-0.156**
	(0.013)	(0.013)	(0.014)	(0.018)	(0.019)	(0.069)	(0.069)
Program 50	-0.001	0.026	0.078*	0.055	0.092	0.206	0.223
	(0.028)	(0.037)	(0.043)	(0.053)	(0.065)	(0.252)	(0.281)
Program 51	0.187***	0.170***	0.168***	0.216***	0.207***	0.790***	0.739***
	(0.041)	(0.038)	(0.034)	(0.047)	(0.047)	(0.177)	(0.168)
Program 54	0.021	0.020	0.028	-0.010	0.003	-0.010	-0.011
	(0.030)	(0.032)	(0.038)	(0.043)	(0.046)	(0.171)	(0.168)
Program 55	-0.050**	0.021	0.013	0.012	-0.002	-0.004	0.004
	(0.020)	(0.023)	(0.026)	(0.031)	(0.036)	(0.114)	(0.100)
Program 56	-0.004	0.001	0.002	-0.009	-0.007	-0.036	-0.041
	(0.012)	(0.012)	(0.014)	(0.015)	(0.018)	(0.060)	(0.060)
Program 57	0.023	0.027*	0.036**	0.028	0.030	0.119	0.116*
	(0.015)	(0.015)	(0.015)	(0.019)	(0.020)	(0.074)	(0.065)
Program 58	-0.037	-0.029	-0.011	0.034	0.044	0.118	0.139
	(0.036)	(0.044)	(0.043)	(0.029)	(0.047)	(0.086)	(0.336)
Program 59	-0.011	-0.026	-0.021	-0.026	-0.016	-0.098	-0.095
	(0.016)	(0.017)	(0.018)	(0.020)	(0.023)	(0.078)	(0.076)
Program 60	0.070*	0.059	0.037	0.082	0.052	0.327	0.335
	(0.042)	(0.048)	(0.052)	(0.052)	(0.058)	(0.208)	(0.232)
Program 61	0.101*	0.079	0.104**	0.092	0.156*	0.352	0.325
	(0.052)	(0.052)	(0.053)	(0.074)	(0.082)	(0.292)	(0.245)
Program 62	0.000	-0.004	-0.006	-0.021	-0.021	-0.077	-0.069
	(0.026)	(0.027)	(0.027)	(0.036)	(0.037)	(0.139)	(0.131)
Program 63	-0.063*	-0.045	-0.084**	-0.058	-0.090**	-0.214	-0.218
	(0.038)	(0.036)	(0.035)	(0.042)	(0.039)	(0.151)	(0.144)
Program 65	-0.107***	-0.100**	-0.105**	-0.107*	-0.104*	-0.392*	-0.415**
	(0.038)	(0.050)	(0.052)	(0.065)	(0.063)	(0.211)	(0.203)
Program 66	-0.017	-0.005	0.039	-0.036	0.004	-0.136	-0.221
	(0.037)	(0.040)	(0.044)	(0.055)	(0.056)	(0.210)	(0.190)
Program 67	0.111***	0.018	-0.017	0.020	-0.034	0.077	0.071
	(0.033)	(0.034)	(0.034)	(0.042)	(0.046)	(0.153)	(0.119)
Program 68	-0.020	-0.012	-0.020	-0.019	-0.026	-0.074	-0.068
	(0.013)	(0.014)	(0.016)	(0.019)	(0.021)	(0.071)	(0.072)
Model	OLS	OLS	OLS	OLS	OLS	OP	GOP

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Outcome	Mean	Mean	Mean	SPR	SPR	SPR	SPR
District FE	N	Y	N	Y	N	N	N
School FE	N	N	Y	N	Y	N	N
Model 1 Correlation	1.000	0.874	0.788	0.765	0.683	0.781	0.771
Model 2 Correlation	0.874	1.000	0.902	0.836	0.757	0.851	0.852
Model 3 Correlation	0.788	0.902	1.000	0.784	0.851	0.794	0.778
Model 4 Correlation	0.765	0.836	0.784	1.000	0.922	0.996	0.986
Model 5 Correlation	0.683	0.757	0.851	0.922	1.000	0.922	0.902
Model 6 Correlation	0.781	0.851	0.794	0.996	0.922	1.000	0.992
Model 7 Correlation	0.771	0.852	0.778	0.986	0.902	0.992	1.000

Note: Table presents alternative specifications of summative performance models. All models include controls for potential teaching experience, school gender, race/ethnicity, subsidized lunch status, Limited English Proficiency, and special education composition. The model row indicates the estimation method. Abbreviations: "OLS" = ordinary least squares, "OP" = ordered probit, "GOP" = generalized ordered probit (Pudney & Shields, 2000). In the ordered probit and generalized ordered probit, we include district means of all of the institution indicators as additional controls. These are additionally entered into the threshold indices in the generalized ordered probit model. The outcome row indicates the outcome used in the model. "Mean" = average rating on four standards, "SPR" = summative performance rating. Correlations in the table indicate the correlation of estimated institution effects with the results from the specified model. Standard errors are clustered by teacher. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Attrition Models

The final research question asks whether there are differences in teacher retention across licensure pathways and educator preparation programs. We estimate the probability that a teacher departs the Massachusetts public school system in a given year. In particular, we estimate duration models of teachers' careers in public schools that include the indicators of teachers' preparation included in Eq. (A.1). We summarize this approach with the following empirical model:

$$\Pr\left(Quit_{jst} = 1 \,\middle|\, T \geq t\right) = f(Exp_{jt}\lambda + X_{jst}\beta + P_{jt}\delta). \quad \text{(A.3)}$$

In Eq. (A.3), *j* indicates teacher, *s* indicates school, and *t* indicates the year of service. We estimate the likelihood that a teacher quits in year *t* given that she has not yet left the school system; that is, her total tenure in Massachusetts public schools (*T)* is at least *t* years long. As before, we include teacher and school characteristics (*X*) and indicators of teacher preparation (*P*). Models like Eq. (A.3) have been used to study the career pathways of teachers with different preparation backgrounds and the influence of a number of school and district characteristics on teacher retention (Clotfelter et al., 2008; Goldhaber & Cowan, 2014; Goldhaber et al., 2011, 2016; Imazeki, 2005; Kane et al. 2008).

The model in Eq. (A.3) implicitly assumes that school-level factors omitted from *X* that affect teacher attrition are not related to educators' preparation background. Because this may not be a reasonable assumption, we follow the approach suggested by Wooldridge (2010) for estimation of nonlinear models with unobserved heterogeneity. We include controls for the school (or school-year) means of each of the variables in the model, including the teacher preparation indicators. The institution indicators are therefore identified by comparing teachers in schools with similar staffing composition, but who completed programs at different preparation providers. In linear models, this specification produces coefficients that are identical to the school (school-year) fixed effects model. Goldhaber and Cowan (2014), who employ similar models in an analysis of providers in Washington state, find that they produce similar estimates as linear probability models using fixed effects. This is also the case in Massachusetts, with correlations of the organization indicators of about 0.92-0.95.

In **Table A.5**, we present results from the full set of models.

Table A.5. Alternative Specifications of Attrition Models

	(1)	(2)	(3)	(4)	(5)	(6)
Program 1	-0.030*	-0.048**	-0.059***	-0.033***	-0.044***	-0.057***
	(0.016)	(0.019)	(0.019)	(0.012)	(0.015)	(0.017)
Program 3	0.026	0.017	0.015	0.022	0.013	0.007
	(0.022)	(0.025)	(0.026)	(0.029)	(0.031)	(0.030)
Program 4	0.001	-0.001	-0.010	-0.007	-0.013	-0.021
	(0.012)	(0.014)	(0.016)	(0.012)	(0.016)	(0.017)
Program 5	-0.009	-0.003	-0.003	-0.017	-0.019	-0.012
	(0.015)	(0.018)	(0.019)	(0.014)	(0.018)	(0.020)
Program 6	0.007	0.009	0.004	-0.001	0.000	-0.004
	(0.009)	(0.010)	(0.011)	(0.009)	(0.012)	(0.013)
Program 7	-0.005	0.003	-0.010	-0.012	-0.002	-0.016
	(0.018)	(0.022)	(0.022)	(0.017)	(0.021)	(0.022)
Program 8	0.002	0.007	0.001	-0.005	-0.005	-0.007
	(0.011)	(0.013)	(0.014)	(0.011)	(0.014)	(0.015)
Program 10	-0.016	-0.017	-0.002	-0.023	-0.018	-0.010
	(0.025)	(0.027)	(0.029)	(0.021)	(0.022)	(0.025)
Program 13	0.036***	0.040***	0.041***	0.035***	0.041***	0.040***
	(0.008)	(0.009)	(0.010)	(0.011)	(0.013)	(0.014)
Program 14	-0.014	0.030	0.004	-0.021	0.019	-0.007
	(0.029)	(0.039)	(0.037)	(0.025)	(0.030)	(0.027)
Program 16	-0.009	-0.029	-0.036	-0.017	-0.027	-0.043
	(0.041)	(0.045)	(0.050)	(0.034)	(0.040)	(0.047)
Program 17	-0.002	-0.046	-0.052	-0.011	-0.058	-0.068
	(0.034)	(0.057)	(0.072)	(0.032)	(0.053)	(0.067)
Program 18	-0.044*	-0.040	-0.017	-0.044***	-0.037**	-0.020

	(1)	(2)	(3)	(4)	(5)	(6)
	(0.023)	(0.026)	(0.027)	(0.014)	(0.017)	(0.019)
Program 21	0.007	-0.005	0.030	-0.000	0.002	0.041
	(0.025)	(0.035)	(0.040)	(0.032)	(0.055)	(0.066)
Program 22	-0.018**	-0.003	0.001	-0.024***	-0.013	-0.008
	(0.008)	(0.010)	(0.011)	(0.007)	(0.010)	(0.011)
Program 23	-0.033*	-0.034	-0.030	-0.035***	-0.039*	-0.034
	(0.019)	(0.025)	(0.025)	(0.013)	(0.020)	(0.022)
Program 24	0.008	0.011	0.008	0.000	0.004	0.000
	(0.007)	(0.009)	(0.009)	(0.007)	(0.010)	(0.010)
Program 25	-0.002	0.003	-0.005	-0.009	-0.002	-0.013
	(0.013)	(0.014)	(0.015)	(0.013)	(0.015)	(0.017)
Program 26	0.002	-0.020	-0.015	-0.006	-0.035	-0.023
	(0.027)	(0.028)	(0.032)	(0.029)	(0.034)	(0.038)
Program 28	0.030***	0.023**	0.020*	0.028**	0.019	0.015
	(0.009)	(0.010)	(0.011)	(0.012)	(0.014)	(0.016)
Program 29	-0.032***	-0.019	-0.020	-0.035***	-0.024**	-0.024*
	(0.012)	(0.015)	(0.015)	(0.008)	(0.012)	(0.013)
Program 30	0.032***	0.017	0.017	0.029*	0.015	0.011
	(0.011)	(0.012)	(0.014)	(0.015)	(0.017)	(0.020)
Program 31	0.080***	0.067**	0.083***	0.113**	0.105**	0.118**
	(0.025)	(0.028)	(0.032)	(0.052)	(0.051)	(0.060)
Program 32	-0.007	-0.014	-0.016	-0.013	-0.035	-0.024
	(0.023)	(0.030)	(0.030)	(0.020)	(0.030)	(0.031)
Program 33	-0.011	0.034*	0.020	-0.017	0.019	0.008
	(0.016)	(0.018)	(0.020)	(0.013)	(0.017)	(0.018)
Program 34	0.027	0.049**	0.044*	0.027	0.054*	0.047
	(0.017)	(0.021)	(0.023)	(0.025)	(0.029)	(0.035)
Program 37	-0.025	-0.035*	-0.034	-0.029**	-0.038**	-0.039*
	(0.017)	(0.021)	(0.022)	(0.014)	(0.019)	(0.021)
Program 39	0.005	0.023	0.033*	-0.002	0.015	0.024
	(0.015)	(0.018)	(0.019)	(0.017)	(0.020)	(0.022)
Program 41	0.032	0.023	0.028	0.030	0.024	0.025
	(0.021)	(0.023)	(0.024)	(0.030)	(0.035)	(0.036)
Program 43	0.071**	0.062*	0.083**	0.091	0.086	0.111
	(0.033)	(0.032)	(0.039)	(0.063)	(0.065)	(0.073)
Program 44	-0.044**	-0.048**	-0.053**	-0.042***	-0.050***	-0.050***
	(0.021)	(0.023)	(0.024)	(0.012)	(0.017)	(0.017)
Program 46	-0.018	-0.028	-0.035	-0.026	-0.035	-0.041

	(1)	(2)	(3)	(4)	(5)	(6)
	(0.018)	(0.021)	(0.023)	(0.016)	(0.022)	(0.025)
Program 48	-0.038	-0.033	-0.047	-0.037	-0.044	-0.045
	(0.041)	(0.043)	(0.047)	(0.027)	(0.035)	(0.037)
Program 49	0.013	0.014	0.017	0.006	0.009	0.011
	(0.008)	(0.010)	(0.011)	(0.010)	(0.012)	(0.013)
Program 50	-0.047	-0.064	-0.074	-0.045	-0.040	-0.064
	(0.056)	(0.058)	(0.054)	(0.033)	(0.036)	(0.041)
Program 51	0.124***	0.129***	0.130***	0.244***	0.269***	0.255***
	(0.018)	(0.019)	(0.023)	(0.049)	(0.050)	(0.055)
Program 54	0.019	0.037	0.030	0.014	0.032	0.026
	(0.020)	(0.023)	(0.026)	(0.025)	(0.030)	(0.035)
Program 55	-0.034***	-0.040***	-0.041**	-0.036***	-0.044***	-0.043***
	(0.013)	(0.015)	(0.017)	(0.009)	(0.014)	(0.016)
Program 56	0.001	0.001	-0.001	-0.007	-0.008	-0.008
	(0.010)	(0.012)	(0.013)	(0.010)	(0.013)	(0.014)
Program 57	-0.028***	-0.015	-0.014	-0.032***	-0.021**	-0.020*
	(0.010)	(0.012)	(0.012)	(0.008)	(0.010)	(0.012)
Program 58	-0.076	-0.062	-0.063	-0.053***	-0.051*	-0.043*
	(0.052)	(0.053)	(0.053)	(0.018)	(0.027)	(0.025)
Program 59	-0.009	0.020	0.019	-0.016*	0.009	0.010
	(0.010)	(0.013)	(0.014)	(0.009)	(0.013)	(0.013)
Program 60	0.042	0.033	0.022	0.048	0.036	0.026
	(0.029)	(0.033)	(0.036)	(0.046)	(0.052)	(0.055)
Program 61	0.005	-0.007	-0.019	-0.002	-0.025	-0.025
	(0.038)	(0.047)	(0.046)	(0.044)	(0.050)	(0.054)
Program 62	0.031**	0.024	0.033*	0.028	0.015	0.028
	(0.014)	(0.016)	(0.018)	(0.020)	(0.024)	(0.027)
Program 63	-0.018	-0.045	-0.033	-0.024	-0.047**	-0.041
	(0.023)	(0.030)	(0.030)	(0.020)	(0.023)	(0.027)
Program 65	-0.052*	-0.063	-0.041	-0.046***	-0.045	-0.036
	(0.027)	(0.044)	(0.044)	(0.015)	(0.032)	(0.036)
Program 66	0.020	0.042	0.053	0.013	0.035	0.047
	(0.027)	(0.029)	(0.033)	(0.033)	(0.039)	(0.045)
Program 67	0.034***	0.032*	0.035*	0.035*	0.041	0.040
	(0.013)	(0.017)	(0.019)	(0.019)	(0.027)	(0.031)
Program 68	-0.032***	-0.034**	-0.038**	-0.037***	-0.043***	-0.043***
	(0.012)	(0.016)	(0.016)	(0.009)	(0.014)	(0.015)
Model	Probit	Probit	Probit	OLS	OLS	OLS

	(1)	(2)	(3)	(4)	(5)	(6)
School CRE/FE	N	Y	N	N	Y	N
School-Year CRE/FE	N	N	Y	N	N	Y
Model 1 Correlation	1.000	0.903	0.901	0.942	0.882	0.880
Model 2 Correlation	0.903	1.000	0.962	0.857	0.921	0.905
Model 3 Correlation	0.901	0.962	1.000	0.861	0.910	0.944
Model 4 Correlation	0.942	0.857	0.861	1.000	0.953	0.943
Model 5 Correlation	0.882	0.921	0.910	0.953	1.000	0.977
Model 6 Correlation	0.880	0.905	0.944	0.943	0.977	1.000

Note: Table presents alternative specifications of attrition models. All models include controls for observed teacher experience, school gender, race/ethnicity, subsidized lunch status, Limited English Proficiency, and special education composition. The model row indicates the estimation method. Abbreviations: “OLS” = ordinary least squares, “Probit” = probit. The school/school-year correlated random effects/fixed effects row indicates whether the model includes correlated random effects (probit) or fixed effects (OLS). The correlated random effects models include school/school-year means of all of the observation-varying variables and are estimated using the heteroskedastic probit formulation suggested by Wooldridge (2010). We model the heteroskedasticity with the log of the number of teachers. Correlations in the table indicate the correlation of estimated institution effects with the results from the specified model. Standard errors are clustered by teacher. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Estimating Program Effects Using Student Growth Percentiles

In our primary analyses, we estimate teacher effects on standardized test scores using the MCAS and PARCC tests with value-added models that adjust for differences in student backgrounds using both prior test scores and student and classroom covariates, such as subsidized lunch status and other program participation (e.g., Boyd et al., 2009; Goldhaber et al., 2013b; Koedel et al., 2015b; Mihaly et al., 2013). There is a substantial literature documenting the properties of teacher effectiveness measures based on these models and the relationship with other indicators of teacher effectiveness.[38] As part of the educator evaluation framework, Massachusetts ESE uses a different method of analyzing student assessment data called *median growth percentiles.* The median growth percentiles calculated by ESE are based on a measure called *student growth percentiles* (SGPs). Student growth percentiles range from 1 to 99 and estimate the percentile rank of a student among those with the same test score history. A student growth percentile of 50 indicates that half of all students with similar achievement histories had higher scores on the current test and half had lower scores. Higher growth percentiles reflect achievement gains that are above the median. An estimate of teacher or program effectiveness can then be estimated by taking the mean (or median) of the individual SGPs.

Despite being measured on a different scale, student growth models are conceptually similar to those we estimate. The most important distinction between the two approaches is in how they specify the model of student achievement. SGPs include controls for as many prior test score results as are available but make no adjustment for student demographics or school or classroom characteristics. The models we estimate include student background measures and controls for classroom composition. There has been some debate in the literature about the optimal design of value-added models. The sparser controls in the SGPs are consistent with non-regulatory guidance from the U.S. Department of Education (2009), which argues against the use of student demographics or school characteristics based on the concern that controlling for student demographics in a growth model implicitly sets different expectations for improvement for each group. Such standards could lead schools to treat students differently based on socioeconomic characteristics. On the other hand, omitting controls for student and school characteristics may disadvantage preparation providers that serve low-income or other hard-to-staff schools.

We assess the consequences of omitting student background and school and classroom composition using SGP measures calculated by ESE. Using the sample of teachers in our value-added samples, we first regress SGPs on teacher experience and institution indicators to form experience-adjusted mean growth percentiles for each institution. We then introduce the additional variables in our value-added models. The student growth percentiles control for students' test score history already; if these controls remove bias due to student and school characteristics, we would expect that adding additional controls to the model would not change the estimates of institution effects significantly. In **Table A.6**, we present the results of this exercise. The estimates in the baseline column are from regressions of student growth percentiles on institution indicators, experience, and school year. In the additional covariates column, we include all of the student, classroom, and school characteristics included in the value-added

[38] See, for instance, Bacher-Hicks et al. (2014), Clotfelter (2006, 2007, 2010), Chetty et al. (2014a), Goldhaber et al. (2013), Guarino et al. (2015a,b), Harris & Sass (2011, 2014), Koedel & Betts (2011), Papay & Kraft (2015), Rockoff (2004), and Rothstein (2010).

model. The "estimated bias" in the final column provides the difference between the two sets of estimates and tests its statistical significance. A positive estimate of the omitted variables bias indicates that the mean SGPs are higher than they would be with additional controls (that is, they are biased upward). A negative estimate indicates the opposite. Note that we expect that the two sets of estimates will diverge in finite samples because they include different sets of variables; that is, it would be surprising if there were no difference in the final columns. We therefore test the statistical significance of the differences as well.

Table A.6. Estimating Institution Value Added using Student Growth Percentiles

	Math			ELA		
Organization	**Baseline**	**Additional Covariates**	**Estimated Bias**	**Baseline**	**Additional Covariates**	**Estimated Bias**
Program 1	-0.762	0.732	-1.494*	-5.426***	-3.975***	-1.451
	(2.000)	(1.812)	(0.841)	(1.559)	(1.488)	(1.076)
Program 3	-2.667	-3.316*	0.649	-2.821*	-4.283**	1.462
	(1.898)	(1.771)	(0.777)	(1.598)	(1.722)	(1.169)
Program 4	-0.239	-1.529	1.290**	-1.033	-2.375**	1.342**
	(1.795)	(1.822)	(0.599)	(1.325)	(1.200)	(0.603)
Program 5	-2.529	-2.089	-0.44	-4.897**	-3.360**	-1.537**
	(2.768)	(2.973)	(0.620)	(2.006)	(1.694)	(0.662)
Program 6	0.685	-0.532	1.217**	1.084	-0.069	1.153**
	(1.442)	(1.331)	(0.512)	(1.314)	(1.223)	(0.528)
Program 7	-1.094	-1.272	0.178	-0.889	-2.092	1.202
	(2.200)	(2.071)	(0.591)	(1.383)	(1.448)	(0.788)
Program 8	-1.777*	-2.701***	0.924*	-0.243	-1.495	1.252***
	(1.027)	(0.912)	(0.508)	(1.154)	(1.042)	(0.455)
Program 10	2.399	1.983	0.416	-1.12	-1.901	0.781
	(2.797)	(2.684)	(0.664)	(2.405)	(2.003)	(0.977)
Program 13	1.661	0.167	1.494***	2.457**	1.273	1.184**
	(1.663)	(1.496)	(0.557)	(1.134)	(1.130)	(0.522)
Program 18	1.202	2.365	-1.163	-5.520**	-4.260*	-1.260*
	(2.970)	(2.523)	(1.139)	(2.306)	(2.220)	(0.713)
Program 21	14.389***	14.801***	-0.412	10.121***	8.956***	1.165
	(2.087)	(2.438)	(1.735)	(2.832)	(2.597)	(1.017)
Program 22	-3.808***	-3.450***	-0.358	0.167	0.675	-0.508
	(1.205)	(1.175)	(0.335)	(1.099)	(1.029)	(0.381)
Program 23	-0.815	-0.776	-0.04	0.778	0.175	0.603
	(1.716)	(1.697)	(0.627)	(1.522)	(1.387)	(0.624)
Program 24	-1.688**	-1.456*	-0.232	-0.981	-0.977	-0.004
	(0.753)	(0.772)	(0.354)	(0.814)	(0.799)	(0.347)
Program 25	3.297*	2.342	0.955*	0.43	-0.298	0.728

	Math			ELA		
Organization	**Baseline**	**Additional Covariates**	**Estimated Bias**	**Baseline**	**Additional Covariates**	**Estimated Bias**
	(1.761)	(1.548)	(0.522)	(2.186)	(1.640)	(1.095)
Program 28	3.412**	1.782	1.630***	1.734	0.493	1.241*
	(1.349)	(1.366)	(0.454)	(2.044)	(2.060)	(0.634)
Program 29	-0.9	-0.547	-0.353	0.06	0.061	-0.001
	(1.393)	(1.306)	(0.519)	(1.094)	(1.041)	(0.508)
Program 30	5.019***	2.519*	2.501***	1.083	-0.191	1.274**
	(1.333)	(1.316)	(0.691)	(1.385)	(1.265)	(0.615)
Program 32	-8.626***	-7.867***	-0.758			
	(2.729)	(2.610)	(0.610)			
Program 33	5.286***	4.713***	0.573	0.248	0.024	0.223
	(1.564)	(1.543)	(0.536)	(1.877)	(1.611)	(0.660)
Program 34	-4.52	-4.558	0.038	-0.328	0.996	-1.324
	(2.964)	(3.152)	(1.089)	(2.642)	(2.459)	(1.442)
Program 37	-2.431	-1.85	-0.581	3.628	4.9	-1.272
	(2.822)	(2.407)	(0.785)	(3.835)	(2.980)	(1.425)
Program 39	-5.390**	-1.98	-3.410***	-4.157*	-1.107	-3.050***
	(2.416)	(1.796)	(1.272)	(2.139)	(1.863)	(0.838)
Program 41				1.936	-0.006	1.942*
				(1.717)	(1.364)	(1.100)
Program 44	0.504	-0.039	0.543	-1.426	-1.856	0.43
	(1.968)	(2.143)	(0.560)	(2.125)	(2.097)	(0.586)
Program 46	-3.145	-0.264	-2.881***	-2.466	0.802	-3.269***
	(2.406)	(2.373)	(0.824)	(2.021)	(1.803)	(0.806)
Program 49	-1.374	-1.788*	0.413	-2.728*	-2.298	-0.43
	(1.086)	(1.084)	(0.446)	(1.490)	(1.504)	(0.580)
Program 51	8.454**	9.401**	-0.947	14.560***	15.545***	-0.985
	(3.685)	(3.704)	(1.036)	(4.055)	(3.841)	(1.096)
Program 54	-0.054	-1.763	1.709*	5.248***	4.597**	0.651
	(3.250)	(2.939)	(0.912)	(1.933)	(1.870)	(0.846)
Program 55	-2.624	-0.99	-1.634**	-1.758	-0.761	-0.996*
	(1.619)	(1.464)	(0.656)	(1.282)	(1.226)	(0.595)
Program 56	1.593	0.497	1.097**	2.148**	0.913	1.235***
	(1.350)	(1.173)	(0.490)	(0.973)	(0.882)	(0.453)
Program 57	-4.757***	-3.290***	-1.467**	-4.507***	-3.109***	-1.398***
	(1.388)	(1.242)	(0.581)	(1.033)	(1.085)	(0.447)
Program 59	-1.162	-0.769	-0.393	-0.514	0.036	-0.551
	(1.667)	(1.578)	(0.627)	(1.010)	(0.882)	(0.443)

	Math			ELA		
Organization	**Baseline**	**Additional Covariates**	**Estimated Bias**	**Baseline**	**Additional Covariates**	**Estimated Bias**
Program 60				1.118	1.869	-0.751
				(1.588)	(1.455)	(0.956)
Program 62	1.336	-1.031	2.367**	4.194*	0.618	3.576**
	(2.583)	(2.353)	(1.177)	(2.189)	(2.602)	(1.412)
Program 63	5.115**	4.641***	0.474	4.136**	4.526***	-0.39
	(2.013)	(1.785)	(1.112)	(1.643)	(1.652)	(0.910)
Program 65	-0.258	1.193	-1.451**	-8.926***	-7.515***	-1.411*
	(2.457)	(2.686)	(0.719)	(2.215)	(1.987)	(0.803)
Program 67	-1.038	-1.712	0.673	-2.193	-3.645	1.452
	(1.740)	(2.030)	(1.089)	(2.870)	(2.600)	(0.970)
Program 68	-2.695**	-1.567	-1.128*	-3.196**	-0.887	-2.309***
	(1.254)	(1.283)	(0.593)	(1.316)	(1.320)	(0.479)
Spearman Correlation			0.872			0.857

Note: Table presents institution indicators from models using student growth percentiles. Estimates in the "baseline" column include controls for experience and school year only. Estimates in the "additional covariates" column additionally include all of the covariates in the value-added models. The "estimated bias" indicates the difference between the full and base model. Standard errors estimated using the method described by Gelbach (2016) and clustered by teacher. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. ELA = English Language Arts.

As the results in Table A.6 demonstrate, there is generally a high level of agreement between the two models. The rank correlations between the baseline and full models are 0.87 in math and 0.86 in ELA, suggesting that the addition of student and school covariates is less consequential for ordering institutions than the addition of school fixed effects in the value-added models. For the most part, SGPs and value added provide similar information about institutions. Nonetheless, there are some statistically significant differences in institution estimates between the two models. In both subjects, 12 of the institution effects differ in a statistically significant way, which is more than we would expect by chance. Of these, four institutions have estimates that differ by two or more growth percentiles, which is about 20% of a standard deviation in the SGP teacher quality distribution.[39] The basic empirical finding is consistent with prior examinations of teacher effects estimated from SGPs (Ehlert et al., 2016), and suggests that the SGP specification may influence estimated teacher effectiveness for at least a subset of institutions.

[39] We estimate the standard deviation of teacher effectiveness using a mixed effects model that controls for experience and year and estimates teacher and classroom (teacher-year) effects. One standard deviation in the teacher effects distribution is 9.7 in math and 7.9 in ELA.

ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

Making Research Relevant

LOCATIONS

Domestic

Washington, D.C.
Atlanta, GA
Austin, TX
Baltimore, MD
Cayce, SC
Chapel Hill, NC
Chicago, IL
Columbus, OH
Frederick, MD
Honolulu, HI
Indianapolis, IN
Metairie, LA
Naperville, IL
New York, NY
Rockville, MD
Sacramento, CA
San Mateo, CA
Waltham, MA

International

Egypt
Honduras
Ivory Coast
Kyrgyzstan
Liberia
Tajikistan
Zambia

March 2017