CEDR Working Paper | WP #2014-2

CEDR
Center for Education Data & Research

Assessing the Relationship Between Teacher Performance on Washington State's ProTeach Portfolio and Student Test Performance

James Cowan
Dan Goldhaber

University of Washington Bothell

The Center for Education Data & Research
University of Washington Bothell
3876 Bridge Way N. Ste. 201 Seattle, WA 98103
(206) 547-5585 cedr@u.washington.edu
www.cedr.us

Suggested citation:
Cowan, J. and Goldhaber, D. (2014). Assessing the Relationship Between Teacher Performance on Washington State's ProTeach Portfolio and Student Test Performance. CEDR Working Paper 2014-2. University of Washington, Seattle, WA.

© 2014 by James Cowan and Dan Goldhaber. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission, provided that full credit, including © notice, is given to the source

You can access other CEDR publications at www.CEDR.us/publications.html

Executive Summary

Background

As part of Washington State's efforts to ensure and improve the quality of the teacher workforce, the Professional Educator Standards Board (PESB), with assistance from Educational Testing Services (ETS), has recently introduced a new, evidence-based assessment of teachers called the ProTeach Portfolio. The development of ProTeach was mandated by the Washington State Legislature in 2007. Specifically, PESB was charged with the task of developing and implementing a uniform assessment that demonstrates teachers' knowledge and skills for producing a positive impact on students' learning. The ProTeach assessment, a key component of the state certification program, seeks both to recognize a high level of achievement in the profession and to "prepare educators who are able to assess their professional growth and achievement in light of their impact on student learning" (Washington Administrative Code, §181-79A-007, 2012).

The Center for Education Data and Research (CEDR) at the University of Washington Bothell has been contracted by PESB to analyze the ProTeach assessment and its alignment with student academic achievement on the WASL. PESB has asked us to answer four questions:

1. Does the proposed ProTeach pass/fail cutline identify teachers who are less effective as measured by student achievement?
2. How much, if any, of the variation in student achievement attributable to teachers is accounted for in the ProTeach composite score?
3. How well do the twelve subcategories of the assessment capture variation in student achievement?
4. Does classroom composition affect teacher performance on the ProTeach assessment?

Analytic Methods

In this report, we use value-added models of student achievement to assess the relationship between teacher performance on the ProTeach assessment and teacher effectiveness in increasing student achievement on standardized exams. One method for assessing the impact of teachers on student learning is a statistical model that estimates their "value-added" to students' achievements on standardized tests. Value-added models (VAMs) have long been

used to measure the correlation between teacher credentials and student achievement on state assessments like the Measures of Student Progress (Cantrell et al., 2008; Clotfelter et al., 2006, 2007; Goldhaber and Anthony, 2007).

Value-added models rely on students' test score growth across consecutive years to estimate the impacts of educational intervention. While test scores are certainly not the only measure of what schools and teachers contribute to student achievement, they certainly are an important measure and one that states have deemded to be important enough to use for school accountability purposes. Moreover, recent research has found that value-added models provide unbiased estimates of teacher and school effects and that value-added estimates predict students' longer-term life outcomes such as college attendance and earnings (Chamberlain, 2013; Chetty et al., 2013a,b; Deming, 2014).

Empirical Findings

We find evidence that teachers passing the ProTeach assessment are more effective than those who fail in reading and fail to find such evidence in math. We estimate that teachers who pass the assessment on the first attempt are about 0.045 - 0.050 student standard deviations more effective than teachers who initially fail the assessment, but these differences are only statistically significant for reading test scores. The magnitude of these findings is similar to the estimated differences in teacher effectiveness associated with having a teacher with about 3 or 4 years of experience rather than a novice teacher, or a teacher who is certified by the National Board for Professional Teaching Standards versus one who is not.

When we examine the composite score (rather than a pass/fail indicator), we generally do not find statistically significant results. We estimate a difference of one standard deviation on the assessment corresponds to a difference of about 0.01 - 0.02 student standard deviations in teacher effectiveness but the differences are not statistically significant. Among the entry scores, we find that Entry 2 and its associated criteria have the strongest relationship with student learning. Our results suggest that the ProTeach assessment could capture more variation in teacher effectiveness by placing more weight on this entry score.

Finally, we find that classroom characteristics explain variation in teacher performance on the ProTeach assessment even when we control for measures of teacher effectiveness. The contribution of classroom characteristics is statistically signifiant in both math and reading. The findings suggest that teachers' scores may be influenced by the students they teach.

Conclusions & Recommendations

The difference in average effectiveness is comparable to other licensing policies, such as NBPTS certification and testing-based certification (Cantrell et al., 2008; Goldhaber, 2007; Goldhaber and Anthony, 2007). The point estimates are, however, similar in math and reading and of a comparable magnitude as other portfolio-based certification assessments (Darling-Hammond et al., 2013; Newton, 2010). Despite this, classification errors based on the ProTeach assessment are relatively common. We estimate that 14-18% of teachers who fail the ProTeach portfolio assessment are actually in the top quintile of the teacher effectiveness distribution based on value-added and that 34-37% of teachers who initially failing the assessment are in the top two quintiles.

Reweighting the ProTeach criterion may improve the predictive power of the ProTeach assessment. The ProTeach assessment does appear to contribute information about teacher quality that is independent of both licensing exam scores and teacher value-added and combining ProTeach with other such measures of teacher effectiveness may reduce the rates of misclassification.

1 Introduction

As part of Washington State's efforts to ensure and improve the quality of the teacher workforce, the Professional Educator Standards Board (PESB), with assistance from Educational Testing Services (ETS), has recently introduced a new, evidence-based assessment of teachers called the ProTeach Portfolio. The development of ProTeach was mandated by the Washington State Legislature in 2007. Specifically, PESB was charged with the task of developing and implementing a uniform assessment that demonstrates teachers' knowledge and skills for producing a positive impact on students' learning. The ProTeach assessment, a key component of the state certification program, seeks both to recognize a high level of achievement in the profession and to "prepare educators who are able to assess their professional growth and achievement in light of their impact on student learning" (Washington Administrative Code, §181-79A-007, 2012).

Since 2010, the ProTeach assessment has served as the key requirement for the second-tier Professional Certificate for teachers in Washington State. The initial Residency Certificate is valid for five years. Teachers must complete and pass the ProTeach assessment before they begin a sixth year of service or lose their credential to teach in the state.[1] ProTeach is therefore a requirement for those who wish to make a career of teaching in public schools in Washington.[2]

The ProTeach portfolio system is part of a broader, national movement towards in-service assessments of teacher practice that rely on teacher reflection and the production of specific evidence of student learning. For instance, the National Board for Professional Teaching Standards, which administers a national program of advanced teacher certification, includes "teachers are responsible for managing and monitoring student performance" and "teachers think systematically about their practice and learn from experience" as two of its five guiding propositions. Connecticut's Beginning Educator Support and Training (BEST) assessment was a similar portfolio-based assessment system (Wilson et al., 2010). The ProTeach assessment includes a similar focus around teacher-directed professional de-

[1] Teachers may obtain a two-year renewal of their Residency Certificate and teachers who fail to pass the assessment may obtain an additional renewal in order to re-submit a portfolio.

[2] The final composite score of ProTeach Portfolio will be used to award teachers Washington's Professional Certification. For teachers who hold a Residency Certificate – the first level of certification for teachers new to the profession in Washington – the Professional Certificate is mandatory. Educators are eligible to begin the ProTeach assessment after teaching for a minimum of 2 years and accumulating at least 1.5 FTE (Full Time Equivalent) and must complete their portfolio by the time they have served 9 years in the profession (i.e., no more than 7 additional years working as a teacher). Teachers who fail to earn the Professional Certificate lose their credential to teach in Washington State.

velopment and responsiveness to the particular needs of individual students.

The ProTeach assessment is a portfolio comprised of three entries aligned with the 3 standards and 12 criteria outlined in the Washington Administrative Code (WAC 181-79A-207). Entry One ("Professional Growth and Contributions") measures a teacher's knowledge and skill in the areas of professional growth and its impact on student learning. Entry Two ("Building a Learning Community") assesses a teacher's analysis and reflection on classroom management, cultural sensitivity in relationships, and involving families and communities in the educational process. Entry Three ("Curriculum, Instruction, and Assessment") asks candidates to demonstrate their understanding and skill of curriculum, instruction, and assessment as well as provide reflective evidence from three focus students. Candidates are expected to demonstrate mastery on the standards outlined in the ProTeach assessments' three components and corresponding criteria.

One method for assessing the impact of teachers on student learning is a statistical model that estimates their "value-added" to students' achievements on standardized tests. Value-added models (VAMs) have long been used to measure the correlation between teacher credentials and student achievement on state assessments like the Measures of Student Progress (Cantrell et al., 2008; Clotfelter et al., 2006, 2007; Goldhaber and Anthony, 2007). Value-added models rely on students' test score growth across consecutive years to estimate the impacts of educational intervention. Recent research has found that such models provide unbiased estimates of teacher and school effects and that they predict longer-term life outcomes such as college attendance and earnings (Chamberlain, 2013; Chetty et al., 2013a,b; Deming, 2014).

As part of a pilot study of the ProTeach Portfolio, the Center for Education Data and Research (CEDR) at the University of Washington Bothell was contracted by PESB to analyze the ProTeach assessment and its alignment with student academic achievement on the Measures of Student Progress (MSP) exam. PESB has asked us to answer four questions: (1) Does the proposed ProTeach pass/fail cutline identify teachers who are less effective as measured by student achievement? (2) How much, if any, of the variation in student achievement attributable to teachers is accounted for in the ProTeach composite score? (3) How well do the twelve subcategories of the assessment capture variation in student achievement? (4) Do classroom characteristics influence teacher performance on the ProTeach assessment?

In the next section, we discuss the ProTeach assessment including some basic information on teacher candidates and performance. We also discuss the literature on value-added

and other similar assessments of teachers. We include a brief discussion of the fundamental problem confronting any teacher certification program, such as that supported by ProTeach: making high-stakes decisions about individual teachers with limited information about their teaching practice. In Section 3 we describe the data used in this report and discuss the analytical methods for assessing the relationship between the ProTeach Portfolio and student achievement. We present the main results in Section 4 and then conclude with a discussion of the policy implications of our findings.

2 Background

2.1 The ProTeach Portfolio Assessment

The ProTeach portfolio assessment is a three-part process that is designed to elicit evidence of teachers' professional accomplishments. PESB has established three general standards and 12 individual criteria for teacher practice. The three entries and associated criteria are listed in Table 1. Each entry is scored separately by expert scorers and responses to each criterion are judged based on a 4-point rubric. Every one of the 12 criteria is given a separate score and the sum of these scores is the teacher's composite score. Teachers with a composite score of 31 out of a possible 48 are awarded the Professional Certificate. The primary purpose of the ProTeach Portfolio assessment is to identify effective teachers.

The first entry ("professional growth and contributions") asks teachers to evaluate their own practice, enumerate goals for improvement, and implement a plan of professional development. Teachers describe how their professional goals are related to PESB's standards. The assessment requires teachers to provide specific evidence that their professional development activities relate to their professional objectives and that they lead to improvements in student learning. Entry 1 additionally focuses on the contributions of teachers to their students' learning environments and their schools' academic environment.

The second entry focuses on teachers' interactions with students and parents. Teachers are asked to describe the learning environment in their community including specific examples of how student diversity affects classroom learning. They are then asked to describe the ways in which they build a learning community in their classrooms and how student diversity informs their teaching practice. In addition, the second entry requests information about teachers' strategies for assessing student work and communicating information about performance to students and parents.

The final entry focuses on instruction and assessment. Teachers are asked to identify three individual students and describe the factors that influence their choice of instructional practices. They then describe their assessment of the student needs, how they drafted and communicated learning goals to the students, the ways in which they varied their practice, and evidence of student improvement.

Since January 1, 2010, teachers issued a Residency Certificate are required to have submitted a ProTeach portfolio or complete a professional certificate program to earn the Professional Certificate. Teachers have three years once they become eligible for the assessment (typically after two years of teaching) in which to complete their portfolio. The first

Table 1: ProTeach Assessment Entries and Criteria

Entry	Criterion
1. Professional growth and contributions	2b. using professional standards and district criteria to assess professional performance and plan and implement appropriate growth activities
	2c. remaining current in subject area(s), theories, practice, research and ethical practice
	3a. advocating for curriculum, instruction and learning environments that meet the diverse needs of each student
	3b. participating collaboratively in school improvement activities and contributing to collegial decision making
2. Building a learning community	1c. using appropriate classroom management principles, processes and practices to foster a safe, positive, student-focused learning environment
	1e. demonstrating cultural sensitivity/competence in teaching and in relationships with students, families and community members
	1g. informing, involving and collaborating with families and community members as partners in each student's educational process, including using information about student achievement and performance
3. Curriculum, instruction, and assessment	1a. using instructional strategies that make learning meaningful and show positive impact on student learning
	1b. using a variety of assessment strategies and data to monitor and improve instruction
	1d. designing and/or adapting a challenging curriculum that is based on the diverse needs of each student
	1f. integrating technology into instruction and assessment
	2a. evaluating the effects of his/her teaching through feedback and reflection

Source: Professional Educator Standards Board, "Washington ProTeach Portfolio Assessment: Standards and Criteria," http://www.waproteach.com/overview/standards_criteria.html

Note: The criteria are divided into 3 standards, which is reflected in the numbering. Standard 1 is "knowledge and skills for effective teaching"; standard 2 is "knowledge and skills for professional development"; and standard 3 is "professional contributions."

teachers submitted ProTeach portfolio assessments in June 2010. The number of teachers submitting portfolios has steadily increased from 186 during the 2010-2011 school year to 1,161 during the 2012-2013 school year.

In Table 2, we describe the characteristics of teachers sitting for the ProTeach assess-

ment since the 2010-2011 school year. For comparison, we also include all other certificated employees included in the S-275 during schools years 2011 - 2013.[3] The observations in Table 2 are at the teacher-year level. Across the 2,043 submissions we can link to teachers in the S-275, the average experience is just over 6 years, compared to 14 among all certificated employees without ProTeach submissions. Given that teachers who have already advanced to the second-tier certificate (either Continuing or Professional) are not required to submit a portfolio, this is not surprising. Similarly, teachers with ProTeach submissions are much less likely to have earned an advanced degree (51%) than teachers without submissions (71%). ProTeach participants also spend more time in classroom instructional activities (0.95 FTE) than non-participants (0.78 FTE). Teachers with submissions have similar demographic characteristics as other teachers.

Table 2: Summary Statistics for Certificated Employees, 2010-2013

	ProTeach Submission	No Submission
Experience	6.16	14.31
	(3.94)	(9.68)
Advanced degree	0.51	0.71
	(0.50)	(0.45)
WEST-B	0.06	0.10
	(0.74)	(0.74)
Male	0.25	0.28
	(0.43)	(0.45)
Asian	0.03	0.03
	(0.18)	(0.16)
Black	0.01	0.02
	(0.09)	(0.12)
Hispanic	0.04	0.03
	(0.19)	(0.17)
Teaching assignment FTE	0.95	0.78
	(0.18)	(0.39)
N	2043	201479

Notes: Summary statistics derived from S-275 and ProTeach assessment files for the 2010-2011 through 2012-2013 school years. Observations in the "All Certificated" column include all teachers with a certificated assignment in the given year. Observations in the "ProTeach" column include teachers with a ProTeach submission in the given year.

Teachers are assessed on each of the 12 criteria listed in Table 1 with a possible 4 points

[3]The S-275 is the official employment reporting system of the Washington Office of the Superintendent of Public Instruction. We describe this database more fully, as well as the data used in the main findings of this report in Section 3.

available for each. Two of the three entries are rated by one assessor and the third, which is randomly chosen, is double-scored with the final score being the average of the two raters' assessments. If the assessed scores on the double-scored entry differ by more than one point, the score is adjudicated by an additional rater. In addition, if a rater determines that any entry is non-score-able, the portfolio is considered incomplete and the teacher does not receive a final score. Of the 2,148 total portfolio submissions since 2010, 240 include non-score-able entries and are incomplete. For teachers with a complete submission, a composite score across the criteria of 31 out of a possible 48 is required to receive a passing score. However, teachers who initially fail may resubmit any entry of their choosing and bank the scores from the other entries.

Across the 1,908 complete ProTeach submissions, the mean score is about 34 points and the standard deviation is approximately 3 points. A total of 221 submissions, or approximately 10% of the completed submissions, failed to receive a passing score. Of those who initially failed the assessment, 109 received a passing score on a subsequent submission. The lowest score on the assessment is 14.5; however, most teachers who fail the assessment receive scores close to the threshold. The median score is 34.5 and the highest score is 42.5 out of 48. For the primary findings of this report, we will refer to standardized scores on the ProTeach assessment rather than scaled scores. The standardized score refers to the distance in standard deviations from the mean assessment score.

2.2 Value-added Estimates, Teacher Practice, and Student Outcomes

With the increasing availability of longitudinal data systems that follow individual students over time, value-added models have become a common method of assessing the contributions of teacher credentials to student learning. The earliest research in the field documented two empirical findings that were common across many methods, states, and years. First, there is wide variation in student achievement attributable to individual teachers (Aaronson et al., 2006; Goldhaber et al., 1999; Hanushek, 1992; Rivkin et al., 2005; Rockoff, 2004). Second, many easily observable teacher attributes, with the exception of early-career experience, do not appear to explain much of the variation in teacher effectiveness (Clotfelter et al., 2007; Goldhaber and Brewer, 1997, 2000). The research on predictors of teacher effectiveness has presented something of a challenge to traditional methods of formal teacher evaluation and compensation, which tend not to rely on degree and experience and which have not historically documented substantial variation in teacher effectiveness (Weisberg et al., 2009).

While formal credentials other than experience may not be predictive of student achievement, another strand of the value-added literature has found connections to different assessments of teacher practice. These include high-quality classroom observational rubrics (Grossman et al., 2010; Kane et al., 2011; Mihaly et al., 2013b; Tyler et al., 2012) and evaluations by principals (Jacob and Lefgren, 2008), mentors (Rockoff et al., 2011), and students (Kane and Staiger, 2011). More recent research has also shown that value-added is related to long-term student outcomes such as educational attainment and earnings (Chamberlain, 2013; Chetty et al., 2013a).

Using a dataset that connects detailed classroom observations to student outcomes, Grossman et al. (2010) and Kane et al. (2011) evaluate the contributions of individual teaching skills to measured student outcomes. Kane et al. (2011) find that observed classroom practices predict student achievement, with classroom management skills appearing to be particularly important in math and questioning and discussion leadership appearing to be important in reading. Grossman et al. (2010) also present clear evidence that classroom practices are important for explaining differences in reading achievement. They find that teaching strategies for approaching textual interpretation are especially predictive of performance on standardized exams.

Perhaps most closely related to the current report are several studies that have examined the relationship between performance on licensure assessments and student achievement. For instance, Goldhaber (2007) finds that national teacher licensure exams are predictive of teachers' later performance in the classroom. Several studies have considered the relationship between the assessment used by the National Board for Professional Teaching Standards and teacher effectiveness. These studies have found mixed evidence on the signalling value of the credential (Cantrell et al., 2008; Clotfelter et al., 2007; Goldhaber and Anthony, 2007; Harris and Sass, 2007). However, while Cantrell et al. (2008) find no difference in average effectiveness by certification status, they do find that the composite assessment score predicts student achievement. There is also evidence that the scores on portfolio assessments used in California and Connecticut are related to teacher effectiveness (Darling-Hammond et al., 2013; Newton, 2010; Wilson et al., 2010). An important lesson of this literature is that a positive relationship between performance assessments and student achievement does not guarantee that there will be differences in average effectiveness for groups of teachers above and below whatever cut point determines proficiency (Cantrell et al., 2008; Goldhaber, 2007).

2.3 Using Imperfect Measures of Teacher Quality for High Stakes Decisions

PESB currently uses the ProTeach assessment to set a minimum standard for effective teaching among experienced teachers in Washington State. We illustrate the certification problem in Figure 1. In each of the graphs, the vertical axis represents actual teacher effectiveness. Here, we intend "teacher effectiveness" to mean PESB's definition of good teaching, which may or may not include student performance on standardized exams. Ideally, PESB would directly observe the preferred dimension of effective teaching and award Professional Certification only to those teachers meeting its minimum standard. The standard is illustrated as the dashed horizontal line in Figure 1(a). Teachers with actual performance above the line meet the state's true standard for effective teaching. PESB would prefer to award certification to teachers above the line and deny certification to those below the line.

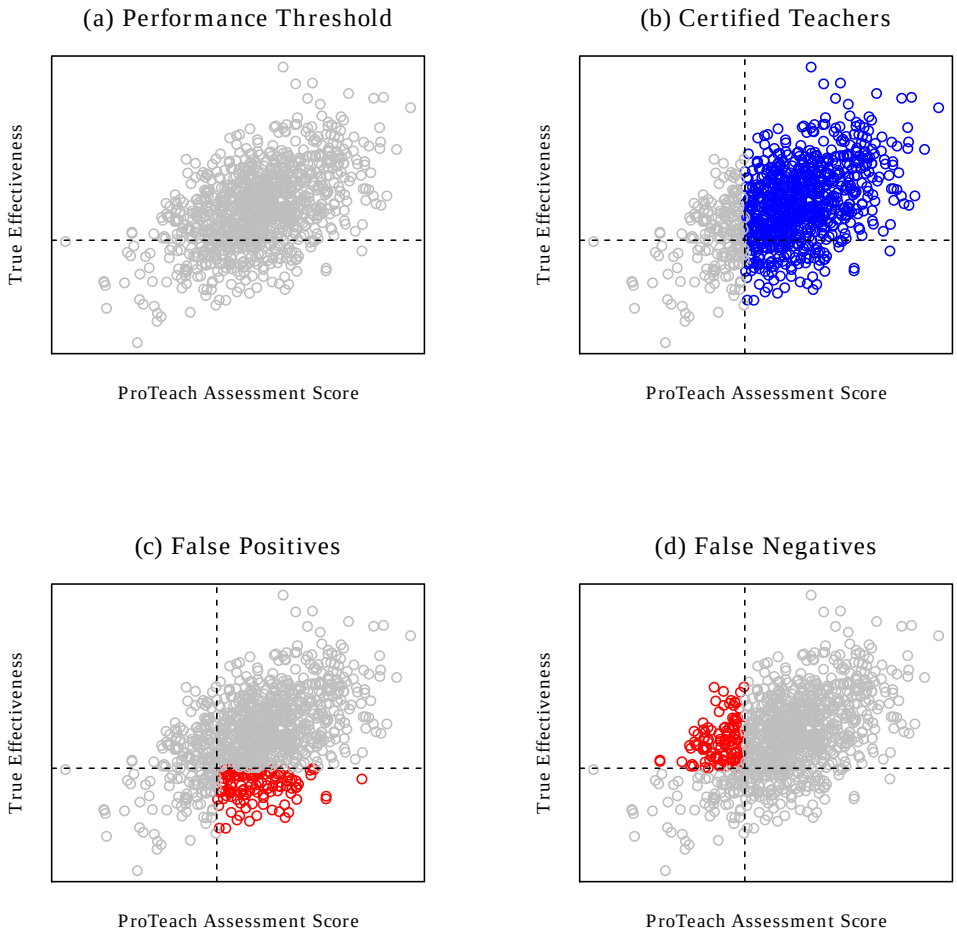
However, we cannot directly know how effective teachers are since this is only experienced by their students and colleagues. Given this, policymakers wishing to use performance to inform licensure or personnel decisions must rely on some measure of teacher effectiveness, such as performance on the ProTeach assessment, which we depict on the horizontal axis in Figure 1. The passing score is indicated by the vertical dashed line in Figure 1(b) and those teachers who receive a passing score are marked in blue.

In the hypothetical data plotted in Figure 1, there is not a perfect relationship between the ProTeach assessment score and actual teacher effectiveness. That is, for any given ProTeach score, there is variation in the effectiveness of teachers with that particular score.[4] Conceptually, there are two reasons the observed ProTeach score diverges from the actual, unobserved teacher effectiveness. First, ProTeach is designed to capture the elements of teaching enumerated in Table 1, which may fail to capture important elements of effective teaching. Second, the ProTeach portfolio encompasses a finite set of assessed activities judged by a small group of raters. Some of the observed variation is likely related to random fluctuations in the performance of individual teachers on particular portions of the portfolio and disagreements among raters about the merit of particular entries rather than to true differences in the performance of teachers.[5] In other words, the preferred measure of

[4] In other words, these measures may be a signal of true teacher effectiveness, but they should not be interpreted as effectiveness itself. For a more in-depth discussion of this issue, see Kane and Staiger (2011, 2012).

[5] These roughly correspond to the notions of validity, which refers to the correlation between a performance measure and true, unmeasured performance, and reliability, which refers to the correlation among subsequent applications of the same performance measure. For more information about these issues in the context of a portfolio assessment system, see Ingvarson and Hattie (2007).

Figure 1: Implications of Using a Performance Measure to Determine Teacher Certification



Notes: In figure (a), the horizontal dashed line represents the desired performance criterion based on actual effectiveness. In figure (b), the vertical dashed line represents the actual certification criterion based on performance on the ProTeach assessment. Teachers marked in blue receive their Professional Certificate. In figure (c), teachers marked in red receive a passing score on the ProTeach assessment but fail to meet the desired performance standard ("false positives"). In figure (d), teachers marked in red receive a failing score on the ProTeach assessment but would exceed the desired performance standard if it were accurately measured ("false negatives").

teacher performance may deviate from actual teacher effectiveness either because it measures skills that do not contribute to teacher effectiveness or because the skills are difficult to consistently measure.

The consequence of the imperfect relationship between actual and measured performance is that decisions about teacher certification are made with limited information about true teacher performance. Therefore, it is not possible to make high-stakes decisions with complete accuracy. There are two kinds of possible mistakes given such a certification policy. On the one hand, some teachers who fail the ProTeach assessment will actually be effective teachers whom PESB would prefer to license if actual effectiveness were known ("false negatives"). On the other hand, some teachers who pass the ProTeach assessment will nonetheless be ineffective teachers whom PESB would prefer not to license ("false positives"). The rates of misclassification (that is, the number of teachers in the red categories) suggest the policy trade-offs to using the ProTeach assessment to support individual licensing decisions. If misclassification rates are high, then basing decisions on such assessments may not lead to demonstrable improvements in student outcomes.[6]

In this report, we consider the validity of the ProTeach assessment. It is important to note that the strength of the relationship between the ProTeach assessment score and teacher effectiveness is limited by the precision of the measurement instrument. For instance, if the portfolio rubric captures important dimensions of teacher skill but the raters frequently disagree about individual teachers, we may not observe a strong relationship with teacher effectiveness. If this were the case, increasing the number of raters assigned to assess each prompt could improve the validity of the assessment. Among the January 2013 submissions, raters reached exact agreement on 62% of the criteria that were assessed by multiple evaluators (Educational Testing Service, 2013). In our analyses of the individual rater data, we find that the correlation between different raters' assessments of the same entry, which is referred to as the inter-rater reliability, ranged from 0.20 - 0.45, depending on the criterion.[7] We further discuss the consequences of rater error for our results in the conclusion. However, the results we present here should be interpreted as encompassing both the design of the ProTeach portfolio itself as well the determinations made by PESB about scoring procedures.

[6]For a discussion of this point in the context of teacher licensure exams, see Goldhaber (2007).

[7]For comparison, we present some results from the Measures of Effective Teaching Project, which considered several different performance measures for one set of teachers across a number of school districts (Mihaly et al., 2013b). The value-added measures displayed year-to-year reliabilities in the range of 0.30 - 0.50 for elementary teachers and 0.45 - 0.85 for middle school teachers, who generally teach more students per year. Value-added measures are generally more stable in math than reading. These findings are in line with the previous literature (Goldhaber et al., 2013b; McCaffrey et al., 2009). The classroom observation measures had reliabilities in the range of 0.25 - 0.70 depending on the rubric. Finally, student surveys had a reliability of 0.40 - 0.50 for elementary teachers and about 0.90 for middle school teachers.

3 Data and Analytic Approach

In order to assess the effectiveness of teachers taking the ProTeach assessment, we analyze student outcomes for elementary and middle school students on the Washington Assessment of Student Learning (WASL) and Measures of Student Progress (MSP) exams in grades 3-8 for school years 2010 - 2012. We describe the data used in this report in more detail in Section 3.1. We then provide an introduction to the value-added methods we use in 3.2.

3.1 Data

The data for this report is derived primarily from administrative databases prepared by Washington State's Office of Superintendent of Public Instruction (OSPI) and the Professional Educator Standards Board (PESB). We derive teacher characteristics from the Washington State S-275 personnel report and student demographics and standardized test performance from the the Comprehensive Education Data and Research System (CEDARS) and the Washington Assessment of Student Learning or Measures of Student Progress (WASL/MSP) databases.

The S-275 contains information from Washington State's personnel-reporting process; it includes a record of all certified employees in school districts and educational service districts (ESDs), their place(s) of employment, and annual compensation levels. It also includes gender, race/ethnicity, highest degree earned, and experience. We link teacher data in the S-275 to information on ProTeach outcomes provided by PESB. We use data on ProTeach submissions from June 2010 through January 2013. In all, there are 2,148 submissions and 1,980 complete submissions representing 1,992 unique teachers. The ProTeach database includes scores on the individual criteria as well as entry and composite scores for all teachers who submitted portfolio entries. In addition, we obtained rater-level data from ETS for the double-scored entries. Of the 2,148 total submissions, we are able to link 2,073 to teacher records in the S-275 during the testing year.

Information on teachers in the S-275 and the Washington State Credentials database can be linked to students via the state's CEDARS and WASL/MSP databases.[8] The MSP is Washington's statewide standardized exam for students in grades 3 - 8. Students are tested

[8]We use both WASL and MSP test scores for this analysis. Our main analyses of the ProTeach assessment use data from the 2009-2010, 2010-2011, and 2011-2012 school years using student post-test results from the MSP. For the 2009-2010 school year, our baseline test score is measured using the prior year WASL score. We also use WASL results from the 2007-2008 and 2008-2009 school years to test the predictive power of the ProTeach composite score in some regressions.

in mathematics and reading in each grade and in writing and science in grades 4 - 7 and 5 - 8, respectively. The exams are tailored to the state's Essential Academic Learning Requirements (EALRs) and are written and reviewed by panels of educators in the state. CEDARS includes information on individual students' backgrounds including gender, race/ethnicity, free or reduced-price lunch, migrant, and homeless statuses; as well as participation in the following programs: home-based learning, learning disabled, gifted/highly capable, limited English proficiency (LEP), and special education. It additionally includes a direct link (a unique course ID within schools) between teachers and students.

We combine the data from the sources described above to create a dataset that links teachers to their students for the 2009-2010 to 2011-12 school years in both math and reading. In this report, we study the relationship between teacher performance on the ProTeach portfolio assessment and student achievement on the WASL/MSP exams. The choice of outcome variable and the requirement of a baseline test score to control for pre-existing differences in student achievement restricts our analysis to student performance in grades 4-8. Because high school students are not tested annually in each subject, we omit high school students from our analysis.[9] Our value-added sample includes 13,428 unique teachers. Of these, 443 teachers submitted entries for the ProTeach assessment and 406 teachers submitted a complete portfolio. Of these teachers, 10,371 total teachers and 349 with ProTeach portfolios are linked to students in math classes and 10,780 teachers and 333 with ProTeach portfolios are linked to students in reading classes.

We present summary statistics for the mathematics value-added sample in Table 3 and for the English/Language Arts value-added sample in Table 4. Classroom demographics are similar for teachers with and without ProTeach portfolio submissions. In the math sample, pupils are slightly more likely to be Asian or Hispanic and to receive free or reduced price meals. They are slightly less likely to be receiving gifted services and slightly more likely to be receiving special education services. The baseline math scores are higher than for those whose teacher submits a complete portfolio but similar to those whose teacher's application is incomplete. Teachers with portfolio submissions are much less likely to teach elementary students than teachers without submissions. In the reading sample, pupils of teachers with

[9] A few recent papers have estimated value-added models using high school data (Aaronson et al., 2006; Clotfelter et al., 2010; Goldhaber et al., 2013c; Jackson, 2012b). However, Goldhaber et al. (2013c) and Jackson (2012b) have found that estimates are sensitive to the specification of the learning process and assumptions about teacher assignment. Given the greater difficulty in solving these issues and the small number of high school teachers with ProTeach assessments in tested classrooms, we do not include such teachers in our analysis.

ProTeach submissions are more likely to receive special education and less likely to receive subsidized lunches. Baseline student test scores are also fairly similar.

When we compare the students of teachers with different outcomes on the ProTeach assessment, we see some generally smaller differences. The baseline student achievement is similar for both teachers who pass and fail the ProTeach assessment. Similarly, program participation rates are similar across these samples. However, we do observe some differences with those teachers whose submission is incomplete. Baseline test scores are higher in both the math and reading samples. In math, teachers with incomplete submissions have more Hispanic students and fewer special education students than teachers with complete submissions. In reading, teachers with incomplete submissions have fewer students receiving gifted, special education, or free or reduced price lunch services.

Table 3: Summary Statistics for Math Value-Added Sample

	No ProTeach	Pass	Fail	Incomplete
Math score	0.025	0.016	0.012	-0.019
	(0.985)	(0.991)	(0.990)	(0.980)
Baseline math score	0.027	-0.002	0.003	0.031
	(0.981)	(1.005)	(1.004)	(0.998)
Baseline reading score	0.020	-0.012	-0.008	0.017
	(0.983)	(1.003)	(0.998)	(0.971)
Female	0.494	0.497	0.496	0.490
	(0.500)	(0.500)	(0.500)	(0.500)
Asian	0.087	0.107	0.104	0.083
	(0.282)	(0.309)	(0.305)	(0.276)
Black	0.043	0.057	0.055	0.041
	(0.204)	(0.233)	(0.228)	(0.198)
Hispanic	0.171	0.164	0.175	0.240
	(0.376)	(0.370)	(0.380)	(0.427)
Gifted	0.057	0.050	0.050	0.050
	(0.231)	(0.218)	(0.218)	(0.218)
Limited English proficient	0.046	0.047	0.048	0.050
	(0.209)	(0.212)	(0.213)	(0.218)
Special education	0.104	0.112	0.107	0.079
	(0.305)	(0.315)	(0.309)	(0.270)
Free/reduced-price lunch	0.436	0.458	0.459	0.465
	(0.496)	(0.498)	(0.498)	(0.499)
Elementary classroom	0.459	0.387	0.379	0.330
	(0.498)	(0.487)	(0.485)	(0.470)
N	730877	16548	2333	2732

Notes: Summary statistics derived from CSRS/CEDARS, S-275, and ProTeach assessment files for the 2009-2010 through 2011-2012 school years. The sample is as described in the text and omits classrooms with a teacher who submits a ProTeach portfolio in the same year.

Table 4: Summary Statistics for Reading Value-Added Sample

	No ProTeach	Pass	Fail	Incomplete
Reading score	0.053	0.047	0.054	0.109
	(0.967)	(0.980)	(0.975)	(0.930)
Baseline math score	0.041	0.048	0.061	0.166
	(0.981)	(0.985)	(0.987)	(0.996)
Baseline reading score	0.053	0.047	0.053	0.106
	(0.965)	(0.968)	(0.963)	(0.917)
Female	0.498	0.493	0.490	0.469
	(0.500)	(0.500)	(0.500)	(0.499)
Asian	0.087	0.093	0.095	0.107
	(0.281)	(0.290)	(0.293)	(0.309)
Black	0.042	0.046	0.045	0.036
	(0.201)	(0.210)	(0.207)	(0.187)
Hispanic	0.173	0.153	0.154	0.162
	(0.378)	(0.360)	(0.361)	(0.369)
Gifted	0.058	0.048	0.045	0.024
	(0.234)	(0.214)	(0.208)	(0.155)
Limited English proficient	0.042	0.039	0.040	0.043
	(0.200)	(0.194)	(0.196)	(0.204)
Special education	0.097	0.109	0.106	0.083
	(0.296)	(0.312)	(0.308)	(0.275)
Free/reduced-price lunch	0.436	0.422	0.414	0.355
	(0.496)	(0.494)	(0.493)	(0.479)
Elementary classroom	0.484	0.497	0.503	0.552
	(0.500)	(0.500)	(0.500)	(0.497)
N	687664	12883	1601	1633

Notes: Summary statistics derived from CSRS/CEDARS, S-275, and ProTeach assessment files for the 2009-2010 through 2011-2012 school years. The sample is as described in the text and omits classrooms with a teacher who submits a ProTeach portfolio in the same year.

3.2 Analytic Models

The models discussed below are designed to answer the four research questions proposed by PESB. We first assess whether the cut score used for the ProTeach assessment identifies more effective teachers. We then estimate the relationship between student achievement and teachers' ProTeach assessment scores. Finally, we consider whether the relationship between performance on the ProTeach assessment and student achievement is differs over student populations.

Although the research questions vary, the analytic methods considered in this report are all grounded in a value-added model of student learning (Todd and Wolpin, 2003).[10] These models examine the relationship between measured student achievement gains and the timing of educational interventions in an attempt to uncover the causal effects of specific interventions on student learning. Such models have been used to assess the influence of a wide range of teacher characteristics on student achievement, including experience (Clotfelter et al., 2007, 2010), licensure status (Goldhaber and Brewer, 2000), and training pathway (Boyd et al., 2006, 2009; Goldhaber et al., 2013a; Kane et al., 2008; Mihaly et al., 2013a), as well as the effects of individual teachers (Aaronson et al., 2006; Goldhaber et al., 1999; Rockoff, 2004; Rivkin et al., 2005).

As described above, the teacher-level contribution to student learning estimated in these models are commonly referred to as "teacher value-added." Value-added models estimate the contribution of teachers and teacher characteristics to student learning by regressing measured student achievement on prior achievement and key academic and demographic variables, such as free or reduced price lunch participation, limited English proficiency status, and participation in special education programs. This procedure removes the variation in teacher assignments and contemporary student achievement that can be explained by these variables. The effect of teacher characteristics is then estimated from remaining variation in student test scores, which is mostly comprised of growth from one year to the next.

Value-added is only one possible measure of teacher effectiveness in the classroom. It does not capture teachers' contributions to student learning that are not assessed by the state standardized exam. Indeed, researchers have found that teachers contribute to other components of their students' lives, such as high school graduation, college enrollment, and non-cognitive skills (Chamberlain, 2013; Jackson, 2012a; Koedel, 2008). Nonetheless,

[10] We describe all of our analytic methods in more detail in Appendix A.

value-added provides useful information about student learning for at lease three important reasons. First, there is much greater variation in the measured effectiveness of teachers by value-added than by observable characteristics of teachers and in other performance measures. Second, evidence suggests that standardized test scores predict long-term student outcomes (Murnane et al., 1995). Third, and perhaps most importantly, recent evidence suggests that students who have high value-added teachers are more likely to graduate and attend college and earn more in the workforce (Chetty et al., 2013a). This effect follows teachers across school settings and holds true even given additional information about family background, such as income, that is generally included in value-added models. Furthermore, the long-term effects are substantial even when the effects of teachers on test scores fade out over time.

We estimate the same basic value-added model in several different ways to answer different questions about the relationship of the ProTeach assessment to student achievement. Each of the regression models is designed to estimate differences in average teacher effectiveness by a particular outcome on the ProTeach assessment. To estimate the average difference in effectiveness between teachers who pass and fail the ProTeach assessment, we estimate models that include controls for each of the binary outcomes. We replace these variables with indicators for having completed the portfolio and the composite score to estimate the average difference in effectiveness associated with a one standard deviation difference in teacher performance on the portfolio. We repeat this model with entry scores and also include interactions with student characteristics to estimate differences by student subgroups. Finally, we also estimate models that include other measures of teacher effectiveness, such as WEST-B scores and teacher value-added, to determine whether the ProTeach score adds additional information about student achievement.

Teachers who initially fail the ProTeach assessment are allowed to resubmit one or more entries for the possibility of obtaining a higher score. Previous research on the NBPTS assessment has found that teachers' decisions to retake assessments may degrade the teacher effectiveness signal embedded in maximum scores (Cantrell et al., 2008).[11] Therefore, we use the outcome on the first ProTeach assessment for our main results. However, because PESB grants the Professional Certification based on the highest score a teacher earns, we also consider results for the highest outcome. While the results are substantively similar for

[11] Because only teachers receiving low scores retake the assessment and because teachers can choose which entries to resubmit, retakes are likely to result in higher scores for low-achieving teachers only. Thus, some of the information about differences in teacher effectiveness embedded in the first composite score is lost.

both outcomes, we do find that in general the first outcome better predicts student achievement.

3.3 Robustness Checks

There are several possible obstacles to estimating the relationship between student achievement and measures of teacher effectiveness. First, if students are matched to teachers based on unobserved factors that are correlated with their achievement (conditional on prior measures of achievement), then estimates of the contributions of individual teachers and of measured teacher characteristics will be biased. Rothstein (2010) argues that such sorting invalidates value-added estimates. The most convincing evidence for the validity of value-added measures comes from two sources. One source is the randomization of students to teachers within schools (Kane and Staiger, 2008; Kane et al., 2013). The MET Project, for instance, assessed teachers using value-added, classroom observations, and student surveys, and then randomly assigned teachers to rosters of students within schools. Analyzing post-randomization student achievement, they found that the evaluation measures estimated before randomization accurately predicted student achievement in math. Analyses based on reading test scores were more ambiguous and imprecisely estimated, but not inconsistent with unbiasedness. Randomization within schools, however, does not rule out the possibility that school-level factors bias value-added measures. Chetty et al. (2013b) address this concern in two ways. First, they find that important socioeconomic variables excluded from value-added models, such as family income, have little effect on estimated teacher value-added. Second, they examine the movement of teachers across schools and grades. If teacher value-added measures are biased by school-level factors, then value-added measures estimated in their prior schools should be less predictive of student achievement in their new schools. Yet, Chetty et al. (2013b) find that the predictive power of value-added across schools and grades is consistent with only a small amount of bias.[12]

We adopt several methods to check for potential biases due to the non-random sorting of students to teachers. First, we test whether students who perform unexpectedly well or poorly are systematically assigned to ProTeach teachers according to their ProTeach status. Specifically, we incldue statistical controls for the ProTeach status of a student's teacher in

[12]Because teachers who receive favorable student assignments in one school or grade may also receive favorable assignments in another, Chetty et al. (2013b) actually predicts changes in student test scores across all students in the school and grade with changes in the average value-added of their teachers that is caused by the movement of teachers. Hence, their results cannot be explained by assignment mechanisms in the receiving school or grade.

the following year. The logic of this test is that if students who are assigned a successful ProTeach applicant systematically outperform expectations, then we should see that students who will be assigned a successful ProTeach teacher in the future do so as well. If the coefficients on future teacher status are large and statistically significant, then the estimated effects of having a ProTeach teacher may be biased (Rothstein, 2010).[13] We also estimate versions of our models with school fixed effects so that comparisons of the effectiveness of successful and unsuccessful ProTeach applicants are only made among teachers within the same schools. This ensures that school-level factors, such as neighborhood effects or a supportive principals, that may be correlated with both student achievement and results on the ProTeach assessment, do not influence our results.

A second concern that is particular to measures of teacher effectiveness, such as the ProTeach assessment, is that both standardized exams and the assessment tool may measure common student or classroom attributes. For instance, Entry 2 ("Building a Learning Community") asks teachers to document their classroom management skills and the involvement of parents in student learning. The results of this assessment may reflect student and parent contributions to student learning that are also reflected in student test scores. In particular, a high score on this entry may reflect a particularly involved group of parents or especially attentive students, both of which may be correlated with those students' end-of-year standardized exam scores. If this were the case, then the ProTeach assessment and teacher value-added measured in that year will be *mechanically* correlated due to the common student and parent factors even if no causal relationship between the teacher contribution to the learning community and student achievement exists. That is, we could find a strong relationship between ProTeach performance and student learning that reflects students' contributions to ProTeach results and not true teacher effectiveness.[14] For this reason, for our primary analyses, we omit classrooms during the academic year in which the teacher submits a ProTeach assessment from all of our analyses.

[13] For example, consider two groups of students, one of which is assigned a fifth grade teacher who passed the ProTeach assessment and one of which is assigned a fifth grade teacher who failed the ProTeach assessment. If there are large differences in average test scores among these students in fourth grade test scores, we might conclude that these groups of students are systematically different and that estimated ProTeach effects are biased by nonrandom student assignment. However, such a test depends on the relationship of test score gains within students over time and cannot conclusively establish bias. Schools may assign students to teachers based on prior performance in ways that do not lead to biased estimates (Chetty et al., 2013b; Goldhaber and Chaplin, 2012).

[14] For a more detailed discussion of this issue, see Chetty et al. (2013b) and Kane et al. (2011).

4 ProTeach Portfolio Results and Teacher Effectiveness

In this section, we present the main empirical results from our analysis of the ProTeach assessment. For our main results, we display the regression coefficients on the variables of interest along with their standard errors. Because our analyses rely on a finite sample of teachers, the results are dependent on the sample chosen. The standard errors provide an estimate of the variability in estimated coefficients we should expect across different samples of teachers and students. For some parameters of interest, we also present p-values, which are the probability of observing a regression coefficient at least as large as we have if there were no ProTeach effect. Therefore, smaller p-values represent greater statistical significance. For each of our results, we use stars to denote statistical significance at the 0.10, 0.05, and 0.01 levels. The standard criterion for statistical significance in the social sciences is 0.05.

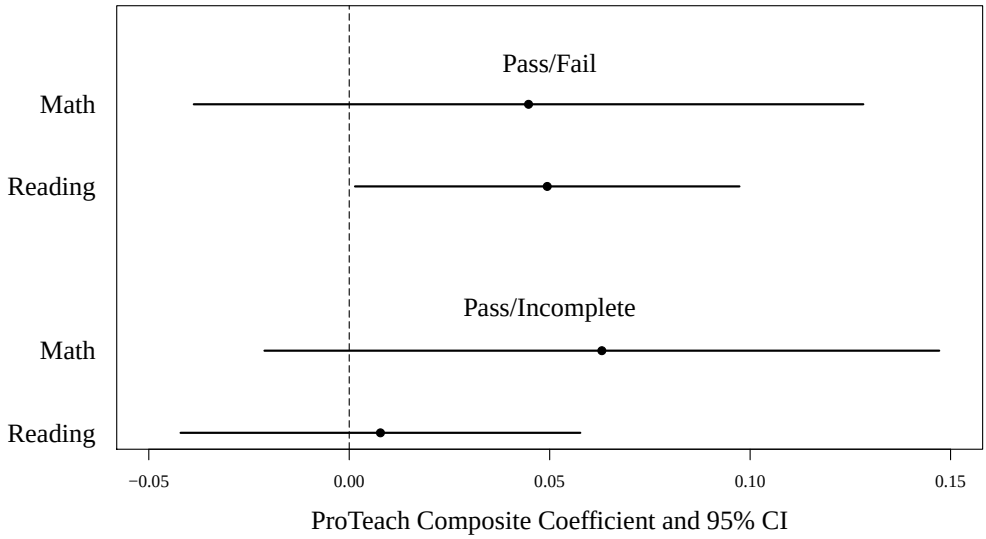
4.1 Professional Certification and Teacher Value-Added

The first research question we address is whether the ProTeach assessment is systematically identifying more effective teachers for professional certification. To begin, we estimate value-added models of student outcomes that include several indicators for teachers' ProTeach assessment status. In particular, we include variables indicating whether a given teacher ever does any of the following: (1) receives a satisfactory score on her ProTeach portfolio submission, (2) receive a failing score on her first ProTeach assessment, or (3) begins a ProTeach submission but initially fail to complete all the assessment activities. These variables are mutually exclusive: teachers can have at most only one of these indicators equal to one. Therefore, these models allow us to compare teachers who passed the ProTeach assessment to those who failed and those who failed to submit a complete portfolio.

We present a graphical summary of these results in Figure 2. The dashed line in Figure 2 indicates no difference in teacher effectiveness. Therefore, points to the right of the dashed line indicate that passing teachers are more effective, while points to the left indicate that passing teachers are less effective. We also plot the 95% confidence intervals as solid lines. Estimated differences for which the solid lines include no effect are not statistically significant.

As the points in Figure 2 demonstrate, we estimate that teachers who pass the ProTeach assessment are more effective than both those who fail and those who do not complete a

Figure 2: Performance on the ProTeach Assessment and Student Achievement



Notes: Figure depicts estimated differences in student achievement by teacher ProTeach status from regressions of student achievement on student characteristics and performance on the ProTeach assessment. "Pass/Fail" indicates average difference in effectiveness between teachers initially passing and initially failing the assessment. "Pass/Incomplete" indicates average difference in student achivement between teachers initially passing and teachers initially submitting an incomplete portfolio. Solid line indicates 95% confidence interval. Student characteristics include: cubic polynomial in lagged math and reading achievement, sex, race, learning disability status, gifted program participation, limited English proficiency participation, special education services participation, free or reduced price lunch participation, classroom means of student variables, and class size. Full regression results are in Table 7.

submission. We estimate that teachers who pass the assessment are 0.045 standard deviations more effective in math than those who fail and 0.050 standard deviations more effetive in reading. However, this difference is only statistically significant for reading achievement; the confidence interval for the math result includes no effect. In other words, we find evidence of an effect of passing the ProTeach assessment on reading achievement, but do not find statistically sgifniciant evidence of an effect for math. Teachers who do not submit a complete portfolio are also less effective than those who pass the assessment although neither of the differences is statistically significant.

The results in Figure 2 are expressed in standard deviations of the student test scores. Depending on the grade level, effects of this magnitude are approximately equal to 1-2 months of learning in reading (Bloom et al., 2008). In terms of teacher credentials, the

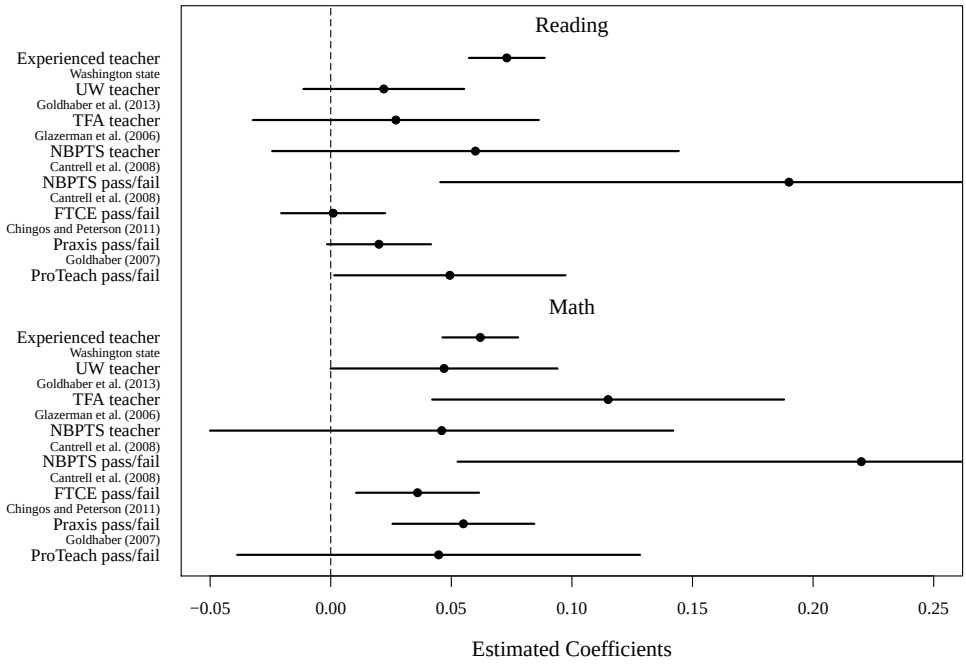
return to the first few years of teaching experience is approximately 0.05 student standard deviations (Clotfelter et al., 2006; Goldhaber et al., 2012; Papay and Kraft, 2013). National Board Certified teachers are about 0.03 - 0.06 standard deviations more effective than other teachers (Cantrell et al., 2008; Goldhaber and Anthony, 2007). The average effectiveness of graduates of Washington State training programs relative to out-of-state teachers ranges from a low of about -0.05 student standard deviations to 0.05 student standard deviations (Goldhaber et al., 2013a).

To provide further context for these results, it is helpful to compare the difference between teachers who pass and fail the ProTeach assessment to other teacher certification methods. Recall that our estimates suggest that teachers who pass the assessment are about 0.05 standard deviations more effective than those who fail. Goldhaber (2007) uses variation in licensing standards over time to examine the relationship between outcomes on the PRAXIS licensing exam and student achievement in North Carolina and finds that the average difference in effectiveness between teachers who would have passed under North Carolina's 2000 standards and those would have failed is 0.07 student standard deviations in math and 0.03 student standard deviations in reading. Chingos and Peterson (2011) compare the effectiveness of teachers who initially pass the Florida teacher certification exam to those who initially fail but subsequently pass and enter the teacher workforce. They find differences of 0.01 to 0.04 student standard deviations between the two groups. One outlier appears to the the National Board for Professional Teaching Standards. Cantrell et al. (2008) exploit the random assignment of NBPTS candidates to classrooms and estimate that successful NBPTS applicants are about 0.20 student standard deviations more effective in both math and reading than unsuccessful applicants. These estimates, however, come from a relatively small sample of teachers and are imprecisely estimated. Other studies relying on larger samples without random assignment have found differences on the order of 0.05 - 0.10 student standard deviations (Cavalluzzo, 2004; Goldhaber and Anthony, 2007). The point estimates in Figure 2 appear to be of the same magnitude of other teacher credentials, particularly those used for assessing candidates for teacher certification.

4.2 Performance on the ProTeach Assessment and Teacher Value-Added

We next consider the relationship between the ProTeach assessment scores and teacher value-added. We display this relationship graphically using measures of teacher value-

Figure 3: Student Achievement Effects of other Teaching Credentials



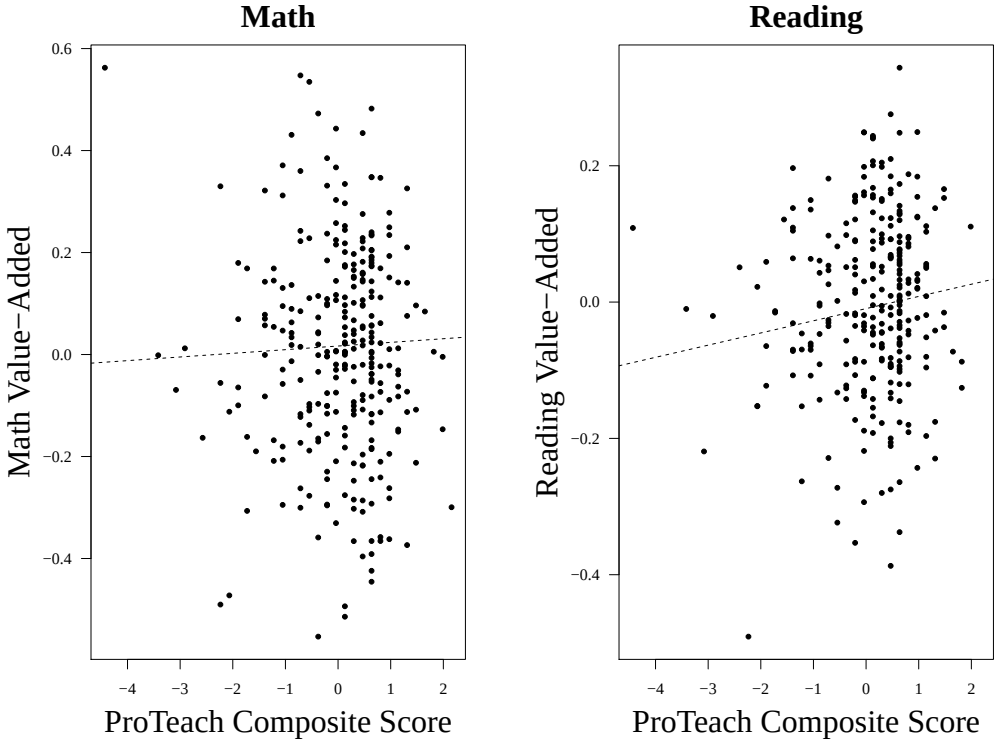
Notes: Figure depicts estimated student achievement effects of selected teacher characteristics and reported 95% confidence intervals. "Experienced teacher" compares novice teacher to those with at least 5 years of teaching experience; "NBPTS" denotes National Board for Professional Teaching Standards; "FTCE" denotes Florida Teacher Certification Exam; "Praxis" is a national teacher certification exam offered by ETS.

added estimated over the same time period.[15] In Figure 4, we plot the ProTeach assessment on the horizontal axis and teacher value-added on the vertical axis. As is evident from the figures, there is wide variation in teacher value-added at each given composite score. That is, for each score on the ProTeach assessment, we tend to observe teachers with a wide range of estimated value-added.

To describe the relationship between the ProTeach composite and teacher value-added more formally, we estimate models that are similar to those that consider ProTeach status. These models include the same controls for student achievement but replace the ProTeach status variables with an indicator for whether the teacher submitted a complete portfolio and the standardized composite score on the portfolio assessment. We plot the regression results

[15]We describe the estimation of the teacher value-added estimates in Appendix A. Because teacher value-added is measured with error in any given year, we estimate the value-added using all available years of data except for the year in which a teacher submitted a ProTeach assessment.

Figure 4: Value-Added and ProTeach Assessment Scores

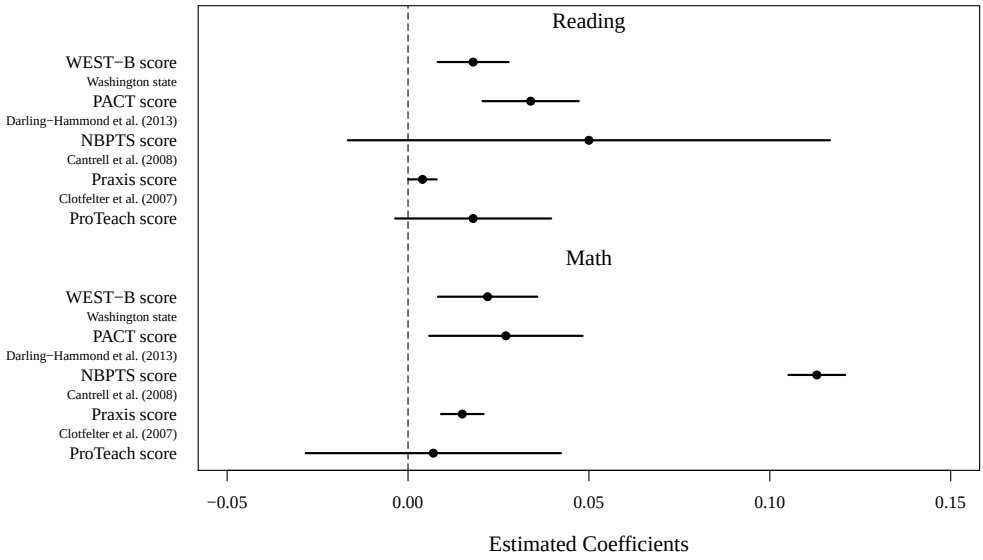


Notes: Standardized ProTeach composite plotted on x-axis. Empirical Bayes teacher value-added across all non-submission years plotted on y-axis. Dashed lines indicate linear fit from regression using student-level data. For math, the estimated slope coefficient is 0.007 ($se = 0.018$). For reading, the estimated slope coefficient is 0.018 ($se = 0.011$).

as the dashed lines in Figure 4. The estimated coefficient on the standardized score is 0.007 ($se = 0.018$) for mathematics. For reading, the coefficient on the composite score is 0.017 ($se = 0.011$). The estimated coefficient measures the difference in effectiveness between two teachers whose ProTeach composite scores differ by 1 standard deviation (or three points). For instance, the coefficient gives the expected difference in teacher effectiveness between a teacher with a ProComp assessment score of 1 standard deviation above the mean and a teacher at the mean. Given the standard errors, we do not find evidence that this difference is statistically different from zero.

For comparison, we present estimates of the effect of one standard deviation on other teacher certification assessments in Figure 5. One standard deviation in teacher value-added in math among Washington State teachers represents approximately 0.18 standard deviations in student learning (Goldhaber et al., 2012). Research based on the NBPTS assess-

Figure 5: Student Achievement Effects of other Teacher Assessments



Notes: Figure depicts estimated student achievement effects of selected teacher characteristics and reported 95% confidence intervals. "PACT" denotes Performance Assessment for California Teachers; "NBPTS" denotes National Board for Professional Teaching Standards; "FTCE" denotes Florida Teacher Certification Exam; "Praxis" is a national teacher certification exam offered by ETS.

ment has found that a one standard deviation difference in the scale score represents a 0.113 difference in student achievement in math and a 0.050 standard deviation difference in reading value-added (Cantrell et al., 2008).[16] Estimates based on certification assessments are more similar to ProTeach. For instance, one standard deviation on teachers' mean WEST-B scores represents a difference of 0.02 standard deviations in both math and reading. Clotfelter et al. (2007) estimate that one standard deviation on the Praxis licensure exam represents about 0.00 - 0.02 student standard deviations.[17] Darling-Hammond et al. (2013) estimate that a one standard deviation difference in assessment results on the Performance Assessment for California Teachers, a portfolio assessment similar to ProTeach, translates to about 0.03 standard deviation in student achievement in both math and reading. Using a smaller sample, Newton (2010) estimates an effect of 0.05 student standard deviations in reading.

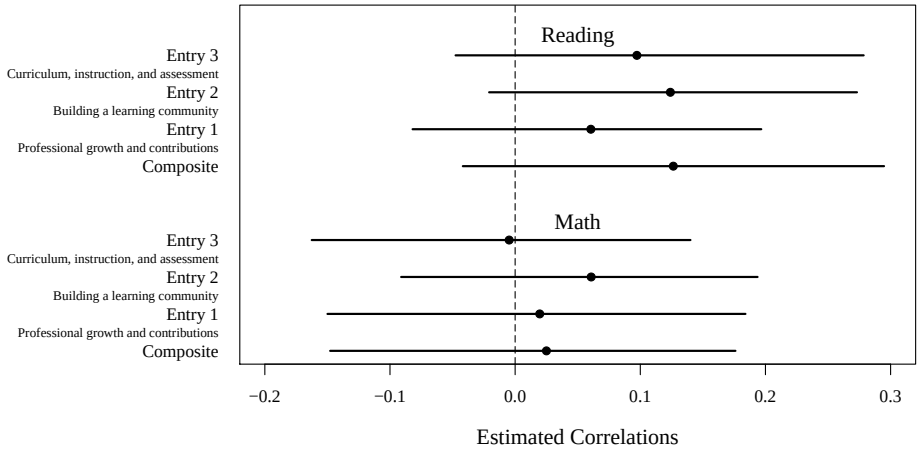
[16]The reading effect was not statistically significant.

[17]Despite the small estimates, the coefficients are statistically significant in both math and reading.

4.3 Performance on Individual ProTeach Criteria and Student Achievement

The ProTeach composite score does not contain all possible information about teacher performance that is included in the assessment. If different parts of the assessment reflect different teaching skills each with different relationships to student achievement, we might find that some portions predict student achievement even when the composite score does not. We begin our assessment of the predictive power of individual components of the ProTeach assessment by disaggregating the composite score into the totals for each of the three entries. We then calculate the correlation between teacher value-added and the entry scores.[18]

Figure 6: Correlation between ProTeach Entry Scores and Teacher Value-Added



Notes: Figure depicts estimated correlations between composite and entry scores on the ProTeach assessment and teacher value-added. The correlations are adjusted for the regression error in the value-added estimates using the method described in Appendix A.

We present the results of this analysis in Figure 6. As in Figure 2, the dots represent our estimates and the solid lines indicate estimated 95% confidence intervals. With the exception of Entry 3 for mathematics, the estimated correlations lie to the right of the dashed line, which inidcates positive estimated correlations between the ProTeach scores and value-added. For the composite score, we estimate correlations of 0.025 in math and 0.13 in reading. We find a weakly positive and statistically insignificant relationship in both

[18]We describe the method used to compute the correlations in the Appendix. The estimation approach corrects the correlations for the sampling variation in the value-added estimates to account for the fact that they are noisy estimates of true teacher effectiveness.

subjects, and a relatively stronger relationship in reading. These findings are consistent with the substantive findings in Figure 4.[19]

The remaining correlations lie between 0 and 0.20 in both math and reading. In both subjects, Entry 2 exhibits the strongest correlation with teacher effectiveness. This correlation is 0.06 for math achievement and 0.12 for reading achievement. The other entry scores have correlations with value-added that are somewhat smaller. We estimate that Entry 1 and teacher value-added have a correlation of 0.02 in math and 0.06 in reading. Our estimates for the correlation of Entry 3 scores and value-added are slightly less than 0 in math and 0.10 in reading. As with the composite score, none of these correlations is statistically significantly different from zero correlation.[20]

Prior research has documented that different teaching skills contribute differently to student achievement (Grossman et al., 2010; Kane et al., 2011). In order to break out the ProTeach results in more detail, we estimate correlations between each of the criterion scores and teacher value-added using the same approach as we did with they entry scores.

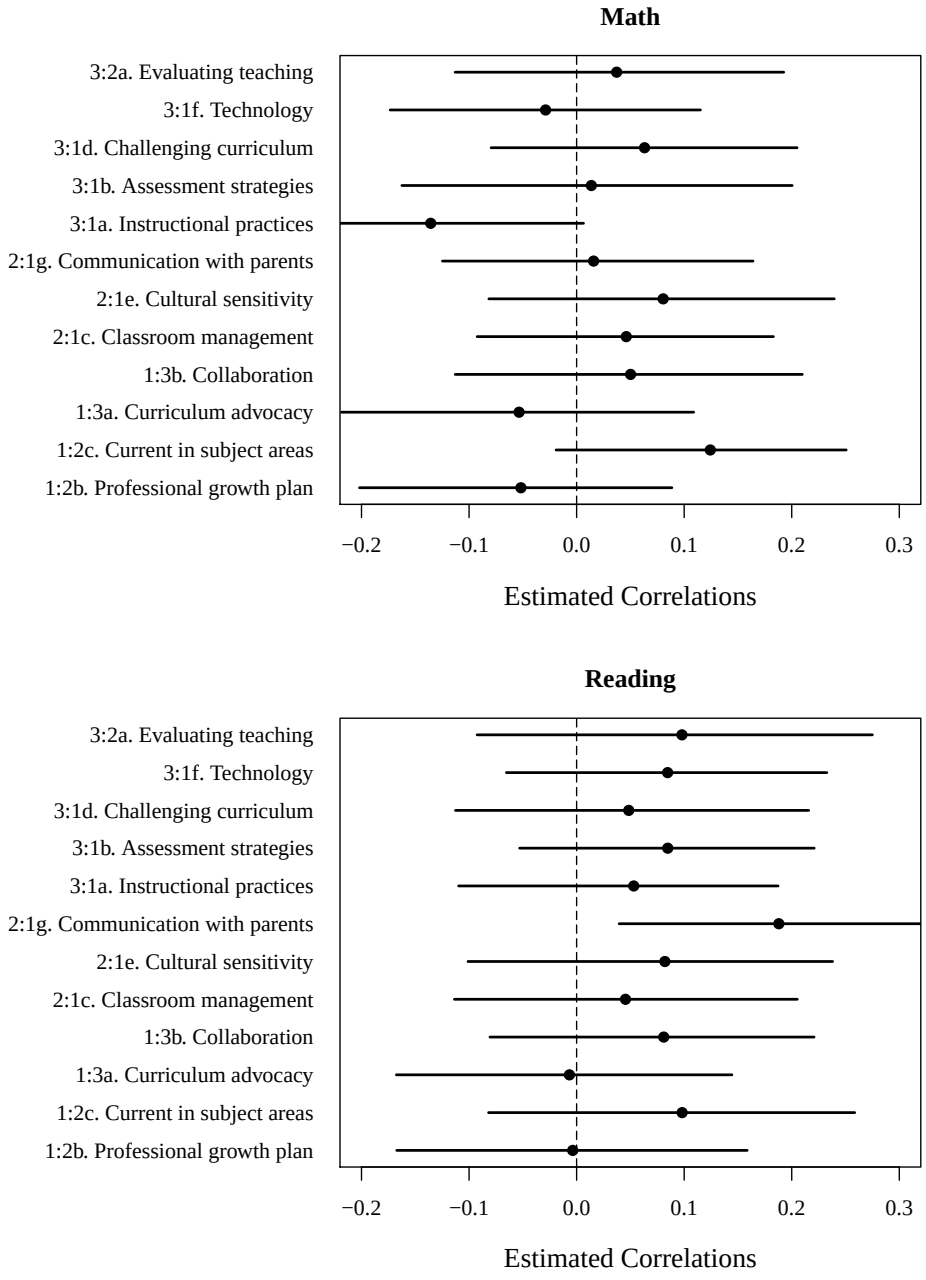
We present the results of this analysis in Figure 7. The magnitudes of the correlations are similar to those of entry scores and value-added with most of the correlations between 0 and 0.20. However, as with the other results, we find that the correlations are mostly statistically insignificant. The single statistically significant result is the correlation between the criterion that measures communication of student performance with parents and reading value-added.

Looking across the criterion scores, the pattern of results mirrors those of the results for entries in Figure 6. For both math and reading, the Entry 2 correlations are uniformly positive. We estimate negative correlations for two of the three Entry 1 criteria in both math and reading, although the estimates are small and statistically insignificant. Reflecting the results for the aggregated entry scores, we observe more mixed results for the Entry 3 criteria. The correlations are generally small or negative in math, while they are positive and somewhat larger in reading. The weights that we estimate in Section 5 reflect the differences in these correlations across criterion scores.

[19]The estimates in Figure 4 are regression coefficients and not correlations. Although the two are related, the point estimates are not directly comparable.

[20]In results not shown, we estimate models like those in Figure 4 using all three entry scores to predict student achievement and test the joint significance of the three entry scores. We fail to reject the null hypothesis that the entry scores are jointly significant in both subjects.

Figure 7: Correlation between ProTeach Criterion Scores and Teacher Value-Added



Notes: Figure depicts estimated correlations between criterion scores on the ProTeach assessment and teacher value-added. The correlations are adjusted for the regression error in the value-added estimates using the method described in Appendix A.

4.4 Classroom Characteristics and Teacher Performance on the ProTeach Assessment

The analyses in the previous section are designed to address whether variation in the ProTeach assessment score has differential effects on particular groups of students. While we find little evidence of differential effects, it may still be the case that teachers who disproportionately teach different populations fare differently on the assessment score. Given the reliance on student voice in the ProTeach portfolio, teachers' scores may depend to some extent on the students they teach independently of any differences in the quality of teacher instruction. For instance, studies relying on surveys of teachers' perceptions of their own effectiveness have found that such measures vary with classroom or school characteristics (Hoy and Woolfolk, 1993; Raudenbush et al., 1992).

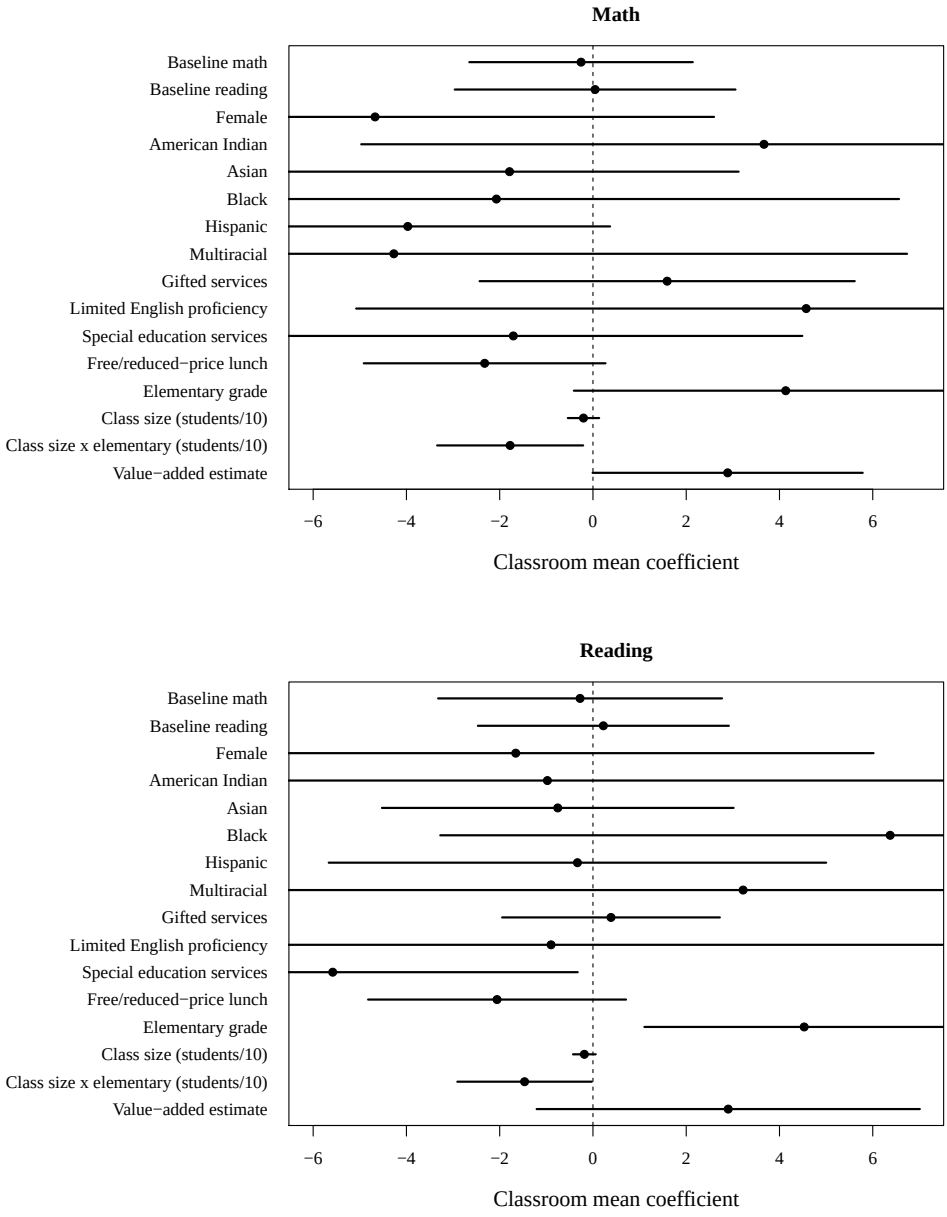
To assess this possibility, we use regression models to predict teachers' performance on the ProTeach assessment using the characteristics of their students in the year in which they took the assessment. If student characteristics are unrelated to teacher performance on the ProTeach assessment, we would expect to find weak relationships between the classroom characteristics and ProTeach results. On the other hand, if student characteristics influence teacher performance, we would expect that adding classroom attributes to a model predicting ProTeach results would add to its explanatory power. We can test this idea formally by testing the *joint significance* of the classroom characteristics.

Comparing the ProTeach performance of teachers with different classroom assignments is complicated, however, by the fact that teachers may not be equally distributed across students. Research suggests that better-credentialed and more-effective teachers are less likely to teach in disadvantaged schools (Clotfelter et al., 2011; Glazerman et al., 2013; Max and Glazerman, 2014; Sass et al., 2012). Therefore, we might see a relationship between teacher performance on the ProTeach assessment and classroom characteristics that is due only to the way in which effective teachers are assigned rather than to an effect of classroom composition on an individual teacher's assessment results.[21] We attempt to control for these assignment patterns by including our value-added measures in these regressions. In other words, we are considering whether classroom characteristics explain variation in ProTeach results that is unrelated to teacher effectiveness.

We present these results in Figure 8. Across math and reading, we find that the number

[21] Conversely, principals might assign more disadvantaged students to more effective teachers. In this case, there might be an effect of student assignments on ProTeach assessment results that is masked by the assignment of more effective teachers to less advantaged students.

Figure 8: Classroom Characteristics and Teacher Performance on the ProTeach Assessment



Notes: Figure depicts estimated coefficients and 95% confidence intervals from regression of ProTeach composite score on classroom mean characteristics, class size, and teacher value-added. The F-statistic from a test of the hypothesis that the coefficients on classroom characteristics are jointly equal to 0 is $F = 1.96$ ($p = 0.02$) for math and $F = 1.99$ ($p = 0.02$) for reading, suggesting that classroom attributes are statistically significantly related to teacher performance on the assessment.

of students in an elementary classroom has a statistically significant and negative effect on teachers' ProComp scores. We also find that elementary teachers tend to perform better on the assessment than middle school teachers. Many of the estimated coefficients are statistically insignificant. That is, the regressions provide little evidence about whether many of the individual student characteristics are related to teacher performance on ProTeach. The pattern of results does suggest that teachers with less advantaged students perform less well on the ProTeach assessment even when we control for teacher value-added. We test this idea formally by testing whether the effects of the classroom and student characteristics on ProTeach scores are jointly equal to zero. For both math and reading, we reject this hypothesis at the 5% level. These findings suggest that classroom composition does influence how teachers perform on the assessment.

Many other measures of teacher proficiency have similar problems. This may be particularly salient for assessments like ProTeach that rely on student voice and evidence of student learning to assess teachers. For instance, Goldhaber et al. (2004) finds that teachers in schools with more free/reduced-price lunch students who apply to the National Board for Professional Teaching Standards are less likely to achieve National Board Certification.

5 Assessing Alternative ProTeach Procedures

The ProTeach assessment comprises both a set of assessment acitivities completed by teachers and procedures determied by PESB for the rating and aggregation of teacher submissions. In order to assess whether those policies influence the relationship between teacher effectiveness and the ProTeach assessment, we consider two components of the PESB scoring procedures. First, we estimate "optimal" weights for the prediction of value-added. We then consider whether the number of raters assigned to each entry affects the relationship between the composite score and teacher effectiveness.

5.1 Optimal Weights for ProTeach Criterion

The current scoring procedure equally weights each of the 12 criterion scores to determine a teacher's total composite score. However, we have seen that some of the entries and criteria have a stronger relationship with teacher value-added than others. Therefore, we might observe a stronger correlation between the two measures of teacher effectiveness by differentially weighting the various components of the ProTeach assessment. We describe the method in more detail in Appendix A, but we choose the weights that best predict our value-added measures.

Table 5: Optimal Weights for Value-Added Prediction

Entry	Criterion	Math	Reading
1. Professional growth and contributions	2b	0.00	0.00
	2c	0.00	0.07
	3a	0.00	0.01
	3b	0.18	0.08
2. Building a learning community	1c	0.00	0.00
	1e	0.70	0.00
	1g	0.07	0.57
3. Curriculum, instruction, and assessment	1a	0.00	0.03
	1b	0.00	0.11
	1d	0.00	0.07
	1f	0.03	0.07
	2a	0.02	0.00

The selection of weights is designed to be the best predictor of teacher effectiveness and not reflect the contribution of individual teacher skills to student achievement. Criteria that capture elements of teaching that are not well measured by other groups of criteria may

therefore receive more weight. Similarly, two highly correlated criterio could each recieve lower weight if they explain a similar facet of teaching.

We therefore begin by estimating weights for each of the aggregated entry scores. As with the results for the correlations with value-added, we find that the optimal weights heavily rely on Entry 2 scores. In fact, for math our optimal weights are 0.14 for Entry 1; 0.86 for Entry 2; and 0.00 for Entry 3. Reflecting the somewhat stronger relationship between Entry 3 and reading value-added, our optimal reading weights are 0.11 for Entry 1; 0.68 for Entry 2; and 0.21 for Entry 3. In both cases, we find that placing substantially more weight on the Entry 2 scores would better predict teacher effectiveness.

The weights are disaggregared by criteria in Table 5. Our estimates suggest substantial variation in the optimal weights across criterion scores with several criteria receiving no weight. As before, the criteria for entry 2, particularly "demonstrating cultural sensitivity/competence in teaching and in relationships with students, families and community members" and "informing, involving and collaborating with families and community members as partners in each student's educational process, including using information about student achievement and performance," receive high weights in both math and reading.

5.2 Rater Reliability and the Relationship between the ProTeach Assessment and Value-Added

While we refer throughout to the relationship between the ProTeach composite score and teacher effectiveness, it is important to note at the outset that the "ProTeach composite" encompasses teacher performance on the portfolio assessment as it is currently constructed. That is, the relationship we observe is necessarily dependent on the selection of assessment items, the particular group of raters, and the policies for rating individual items. Each of these facets of the assessment system introduce additional variation into an individual teacher's assessment score. Importantly, we should not expect the random fluctuations due to, for instance, having had different raters to be correlated with teacher effectiveness.

We can assess the degree to which these random differences in rater judgement influence our result by estimating the rater reliability of the ProTeach composite. Suppose we could submit the same teacher assessment to two entirely separate groups of raters and aggregated the ProTeach assessment scores into a single composite. The reliability measures are intended to estimate the expected correlation between the two scores for each teacher. There are two benefits of this approach. First, it yields some sense of how much varia-

tion in the ProTeach composite represents actual variation in assessed teacher performance rather than differences in rater judgment. Second, we can simulate how the reliability would change under alternative scoring procedures.

Table 6: Reliability of ProTeach Composite under Alternative Scoring Procedures

Procedure	Reliability
All entries single-scored	0.60
Entry 1 double-scored	0.65
Entry 2 double-scored	0.63
Entry 3 double-scored	0.67
All entries double-scored	0.75

We present the results of this analysis in Table 6. Under the current scoring procedures, each entry is graded by a single rater. One of a teacher's entries is randomly chosen to be assessed by a second rater. If the difference between any of the criterion scores across the two raters differs by more than one, the criterion is adjudicated by a third rater. Therefore, every teacher has one entry rated twice and two entries rated a single time. Because the number of criteria varies by entry, the overall reliability of the assessment differs depending on which entry is double-scored. Under the current scoring procedure, we estimate composite reliabilities of 0.63-0.67 depending on which entry is double-scored. This compares to a reliability of 0.60 if all entries were single-scored and 0.75 if all entries were double-scored.

If the rater errors are unrelated to actual teacher effectiveness we can estimate how the rater reliability of the ProTeach assessment influences our results for the relationship between the ProTeach composite and value-added. Because the attenuation in this relationship is inversely proportional to the reliability, these estimates suggest that the estimated coefficients on the standardized composite in Figure 4 would be approximately 50-60% greater if there were no rater disagreements. Of course, this is an impossible ideal: the only way to eliminate rater error is to submit the assessment to an infinite number of raters. A more policy relevant comparison may be comparing the current structure to a scenario in which each entry is double-scored. Our estimates suggest that this would increase the estimated relationship between the composite and teacher value-added by 10-20%.

6 Conclusion and Policy Implications

The Washington ProTeach assessment recommends for Professional Certification teachers who have demonstrated evidence of effective teaching. In this study, we test the validity of the ProTeach assessment against value-added measures of teacher quality. We find a positive relationship between performance on the ProTeach assessment, measured by whether the candidate passes or their scaled score, and student achievement. However, this relationship is generally not statistically significant at conventional levels.

To put these numbers in context, we return to the misclassification model presented above. Recall that the ProTeach assessment can misclassify teachers in one of two ways: teachers who actually fail to meet the state standards can be mistakenly classified as proficient ("false positives") and teachers who do meet the state standards can be mistakenly classified as insufficiently qualified ("false negatives"). We assess the likelihood of misclassification using our value-added estimates.

In order to assess the likelihood of misclassification, we compute the value-added rank of each teacher with a residency certificate in Washington State.[22] We then estimate the probability that a teacher who passes or fails the ProTeach assessment is in each quintile of the teacher effectiveness distribution.[23] We present the results of this simulation in Figure 9. In the Figure, we plot the ProTeach composite score along the horizontal axis and the percentile rank of teacher value-added along the vertical axis. The points in the plot represent individual teachers who took the ProTeach assessment with passing teachers represented as dark circles and failing teachers as empty circles. The passing score (31) is represented as a dashed line.

As is evident from the figure, many teachers who fail the ProTeach assessment have relatively high value-added scores and many who pass have relatively low value-added scores. The shaded rectangles formalize this finding by estimating the probabilit that a teacher who passes (or fails) has actual value-added in each 20% group. As the ProTeach assessment is a noisy signal of teacher quality, classification errors are fairly common. For math, we find that find that 18% of teachers who fail the assessment are actually among the top 20% of teachers by value-added. Among reading teachers, 14% of teachers failing the ProTeach assessment are in the top 20%.

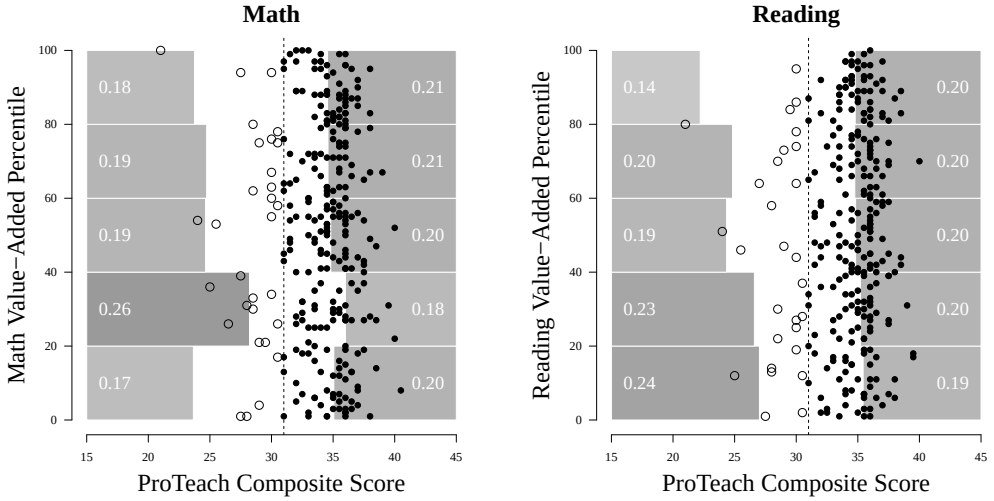
However, the results also indicate that teachers failing the assessment are more likely

[22] Unlike earlier estimates of teacher effectiveness, these are adjusted for differences in experience as this varies considerably in this sample of early-career teachers.

[23] The estimation approach is explained in Appendix A.

to be found among the teachers with the lowest value-added. 43% of teachers failing the assessment are in the bottom 40% of teachers by math value-added compared to 38% of teachers passing the assessment. 48% of teachers failing the assessment are in the bottom 40% of reading value-added compared to 39% among teachers who pass the assessment.

Figure 9: Value-Added and Teacher Classifications



Notes: The x-axis plots the scaled ProTeach composite score and the y-axis plots the percentile of value-added among all teachers with a Residency certificate controlling for experience. Empty circles indicate failing ProTeach scores and filled circles indicate passing ProTeach scores. Numbers in white and shading indicate the probability of having value-added in the given decile of the teacher effectiveness distribution given a failing (left) or passing (right) ProTeach composite score. The probabilities are calculated using the method described in Appendix A.

States and school districts currently use several different methods of evaluating teachers and PESB potentially has many signals of teacher quality available when making certification decisions. It may therefore be of interest whether other evaluation measures have similar misclassification rates. Goldhaber (2006) analyzes the effectiveness of successful and unsuccessful NBPTS applicants. He finds that 55-60% of successful applicants have value-added above the mean of the effectiveness distribution while only 40-45% of unsuccessful applicants are above the mean. Goldhaber and Hansen (2008, 2010) assesses how early-career value-added measures predict later-career value-added measures. They find that 68% of teachers in the bottom quintile of math value-added during the first three years of teaching are in the bottom 40% of value-added in all subsequent years. Among those same teachers, 4% are in the top 20% of value-added in the following years. For reading, 56% of teachers in the bottom quintile of teacher value-added during the first three years

are in the bottom 40% in the years following tenure with 12% in the top 20% of teachers following tenure. Finally, Goldhaber (2007) estimates that about 35% of teachers failing the Praxis certification exam under North Carolina's licensure requirement are more effective than the average teacher.

6.1 Conclusion

In this report, we analyze the predictive validity of the ProTeach assessment using student achievement data from 2010-2012. We estimate that teachers who pass the assessment on the first attempt are about 0.045-0.050 student standard deviations more effective than teachers who initially fail the assessment, although the difference is only statistically significant for reading. The difference in average effectiveness is comparable to other licensing policies, such as NBPTS certification and testing-based certification (Cantrell et al., 2008; Goldhaber, 2007; Goldhaber and Anthony, 2007). When we use the composite score on the ProTeach assessment, we do not find statistically significant results. The point estimates are, however, similar in math and reading and of a comparable magnitude as other portfolio-based certification assessments (Darling-Hammond et al., 2013; Newton, 2010).

Nonetheless, classification errors based on the ProTeach assessment are relatively common. We estimate that 14 - 18% of teachers who fail the ProTeach portfolio assessment are actually in the top quintile of the teacher effectiveness distribution based on value-added. However, we find that reweighting the ProTeach criterion may improve the predictive power of the ProTeach assessment. The ProTeach assessment does appear to contribute information about teacher quality that is independent of both licensing exam scores and teacher value-added and combining ProTeach with other such measures of teacher effectiveness may reduce the rates of misclassification.

References

Aaronson, D., Barrow, L., and Sander, W. (2006). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics*, 25(1):95–135.

Bloom, H. S., Hill, C. J., Black, A. R., and 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., Lankford, H., Loeb, S., and 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., and Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, 31(4):416–440.

Cantrell, S., Fullerton, J., Kane, T. J., and Staiger, D. O. (2008). National board certification and teacher effectiveness: Evidence from a random assignment experiment. *National Bureau of Economic Research Working Paper Series*, 14608.

Cavalluzzo, L. (2004). Is national board certification an effective signal of teacher quality? Technical Report IPR 11204.

Chamberlain, G. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences*, 110(43):17176–17182.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2013a). Measuring the impact of teachers II: teacher value-added and student outcomes in adulthood. Technical Report 19424, National Bureau of Economic Research, Cambridge, MA.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2013b). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. Technical Report 19423, National Bureau of Economic Research, Cambridge, MA.

Chingos, M. M. and Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, 30(3):449–465.

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

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2011). Teacher mobility, school segregation, and pay-based policies to level the playing field. *Education Finance and Policy*, 6(3):399–438.

Darling-Hammond, L., Newton, S. P., and Chung Wei, R. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability*, 25(3):179–204.

Deming, D. J. (2014). Using school lotteries to test measures of school effectiveness. Technical Report 19803, National Bureau of Economic Research, Cambridge, MA.

Educational Testing Service (2013). Washington ProTeach portfolio assessment statistical analysis report. Technical report, Educational Testing Service.

Glazerman, S., Protik, A., Teh, B.-r., Bruch, J., and Max, J. (2013). Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment. Technical Report 2014-4003, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Goldhaber, D. (2006). National board teachers are more effective, but are they in the classrooms where they're needed the most? *Education Finance and Policy*, 1(3):372–382.

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

Goldhaber, D. and Anthony, E. (2007). Can teacher quality be effectively assessed? national board certification as a signal of effective teaching. *Review of Economics and Statistics*, 89(1):134–150.

Goldhaber, D. and 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. and Chaplin, D. (2012). Assessing the "Rothstein falsification test": Does it really show teacher value-added models are biased. *CEDR*, pages 1–39.

Goldhaber, D. and Hansen, M. (2008). Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. Technical Report 3.

Goldhaber, D. and Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, 100(2):250–255.

Goldhaber, D., Liddle, S., and Theobald, R. (2013a). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, 34(1):29–44.

Goldhaber, D., Liddle, S., Theobald, R., and Walch, J. (2012). Teacher effectiveness and the achievement of washington's students in mathematics. *WERA Educational Journal*, 4(2):6–12.

Goldhaber, D., Perry, D., and Anthony, E. (2004). The national board for professional teaching standards (NBPTS) process: Who applies and what factors are associated with NBPTS certification? *Educational Evaluation and Policy Analysis*, 26(4):259–280.

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

Goldhaber, D. D. and Brewer, D. J. (1997). Why don't schools and teachers seem to matter? assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, 32(3):505–523.

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

Goldhaber, D. D., Goldschmitt, P., and Tseng, F. (2013c). Teacher value-added at the high school level: Different models, different answers? *Educational Evaluation and Policy Analysis*, 35(2):220–236.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., and Lankford, H. (2010). Measure for measure: The relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *National Bureau of Economic Research Working Paper Series*, 16015.

Hanushek, E. A. (1992). The trade-off between child quantity and child quality. *Journal of Political Economy*, 100(1):84–117.

Harris, D. N. and Sass, T. R. (2007). The effects of NBPTS-certified teachers on student achievement. *CALDER Working Paper*, 4:1–59.

Hoy, W. K. and Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *The Elementary School Journal*, 93(4):355–372.

Ingvarson, L. and Hattie, J., editors (2007). *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards*. Number 11 in Advances in Program Evaluation. JAI Press, Bingley, UK.

Jackson, C. K. (2012a). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in north carolina. *National Bureau of Economic Research Working Paper Series*, 18624.

Jackson, C. K. (2012b). Teacher quality at the high-school level: The importance of accounting for tracks. *National Bureau of Economic Research Working Paper Series*, 17722.

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

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

Kane, T. J., Rockoff, J. E., and 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. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *National Bureau of Economic Research Working Paper Series*, 14607.

Kane, T. J. and Staiger, D. O. (2011). Learning about teaching. Technical report, Bill and Melinda Gates Foundation, Seattle, WA.

Kane, T. J. and Staiger, D. O. (2012). Gathering feedback for teaching. Technical report, Bill and Melinda Gates Foundation, Seattle, WA.

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

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, 64(3):560–572.

Max, J. and Glazerman, S. (2014). Do disadvantaged students get less effective teaching? key findings from recent institute of education sciences studies. Technical Report NCEE 2014-4010, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., and Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4):572–606.

Mihaly, K., McCaffrey, D. F., Sass, T. R., and Lockwood, J. R. (2013a). 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).

Mihaly, K., McCaffrey, D. F., Staiger, D. O., and Lockwood, J. R. (2013b). A composite estimator of effective teaching. Technical report, Bill and Melinda Gates Foundation, Seattle, WA.

Murnane, R. J., Willett, J. B., and Levy, F. (1995). The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics*, 77(2):251266.

Newton, S. P. (2010). Preservice performance assessment and teacher early career effectiveness: Preliminary findings on the performance assessment for california teachers. Technical report, Stanford Center for Assessment, Learning, and Equity, Stanford, CA.

Papay, J. P. and Kraft, M. A. (2013). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career growth.

Raudenbush, S. W., Rowan, B., and Cheong, Y. F. (1992). Contextual effects of the self-perceived efficacy of high school teachers. *Sociology of Education*, 65(2):150–167.

Rivkin, S. G., Hanushek, E. A., and 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., Jacob, B. A., Kane, T. J., and Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, 6:43–74.

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., Hannaway, J., Xu, Z., Figlio, D., and Feng, L. (2012). Value added of teachers in high-poverty and lower-poverty schools. *Journal of Urban Economics*, 72:104–122.

Todd, P. E. and Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113:F3–F33.

Tyler, J. H., Jacob, B. A., Dougherty, S. M., Hanson, H. J., Fullerton, J. B., and Herlihy, C. M. (2012). Are practice-based teacher evaluations and teacher effectiveness linked in TNTP's performance assessment system ? Technical report.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009). *The Widget Effect*. TNTP, New York.

Wilson, M., Hallam, P., Pecheone, R., and Pamela, M. (2010). Using student achievement test scores as evidence of external validity for indicators of teacher quality: Connecticut's beginning educator support and training program. Technical report, Stanford Center for Opportunty Policy in Education, Palo Alto, CA.

A Analytic Models

A.1 Estimating Teacher Value-Added

Our teacher value-added estimates are estimated from a regression of student test scores on student and classroom characteristics and teacher fixed effects:

$$A_{ijt} = \rho A_{ijt-1} + X_{ijt}\beta + \tau_j + \epsilon_{ijt}. \tag{1}$$

When estimating teacher value-added, we omit years in which a teacher has submitted a ProTeach portfolio.

We then calculate empirical Bayes estimates of teacher value-added using the method of Aaronson et al. (2006). This requires estimating the variance of true teacher effectiveness, which we estimate as the variance of the estimated teacher effects minus the average squared standard error of the estimated effects:

$$\sigma_\tau^2 = J^{-1} \sum_{j=1}^{J} [\tau_j^2 - \sigma^2(\tau_j)].$$

We then estimate the individual teacher value-added as the product of the estimated teacher fixed effect and the estimated reliability ratio of the teacher effect:

$$\tau_j^{eb} = \tau_j \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2(\tau_j)}.$$

This method shrinks teacher effects back toward the mean of the teacher effectiveness distribution in relation to the precision of the estimate. That is, estimates based on more student observations are closer to the estimated teacher fixed effect while those with fewer observations are weighted toward the mean.

A.2 Estimating the Effectiveness of Teachers by ProTeach Status

We begin by analyzing the differences in academic performance of students by the ProTeach status of their classroom teachers. We estimate regressions of student achievement using baseline test scores and student characteristics:

$$A_{ijt} = \rho A_{ijt-1} + X_{ijt}\beta + \delta ProTeach_{jg} + \epsilon_{ijt}. \tag{2}$$

We begin by estimating whether students who have teachers with a ProTeach certification perform better than students of teachers who have failed the ProTeach assessment. In (2), $ProTeach_{jg}$ includes the following indicator variables for teacher j:

1. $Pass_{jg} = 1$ if teacher j receives a passing score on the ProTeach portfolio
2. $Fail_{jg} = 1$ if teacher j receives a failing score on the ProTeach portfolio
3. $Incomplete_{jg} = 1$ if teacher j submits a partial ProTeach portfolio.

The coefficient on $Pass_{jg}$ gives the difference in average achievement gains for students whose teacher completed and passed the ProTeach assessment relative to those students whose teacher has never submitted a portfolio. The remaining coefficients have a similar interpretation. The difference between teachers who pass the assessment and those who fail is

$$\Delta = \delta_{pass} - \delta_{fail}.$$

This difference is our estimate of the signal value of the Professional Certification.

To test whether students assigned to teachers with ProTeach results are different along unobservable dimensions, we estimate regressions that include the ProTeach status of the students' teacher in the following year:

$$A_{ijt} = \rho A_{ijt-1} + X_{ijt}\beta + \delta_1 ProTeach_{jg} + \delta_2 ProTeach_{j,g+1} + \epsilon_{ijt}. \tag{3}$$

If next year's teacher assignment is correlated with the current year test score, it is possible that estimates of ProTeach effects are biased (Chetty et al., 2013b; Goldhaber and Chaplin, 2012; Rothstein, 2010).

A.3 Student Achievement and Teacher Performance on the ProTeach Criteria

To assess the predictive validity of the ProTeach assessment, we replace the indicator for passing the ProTeach assessment in (2) with the teacher's ProTeach assessment score:

$$A_{ijt} = \rho A_{ijt-1} + X_{ijt}\beta + \delta Score_j + \epsilon_{ijt}. \tag{4}$$

The coefficient δ then indicates the average difference in student achievement associated with a one standard deviation increase in the scoring of teacher's ProTeach portfolio.

In addition to an analysis of the ProTeach composite score, we also consider the relationship between the individual ProTeach criterion scores and student achievement. Because of the high level of correlation between the individual ProTeach assessment entries and criteria, we estimate regressions of the form (4) entering each of the entries or criteria separately.

Our analysis of the weights attached to the ProTeach criteria is based on a regression using the the teacher value-added and individual criterion scores. We minimize the mean squared error over sets of feasible weights:

$$\begin{array}{rl} \min & \sum_{i=1}^{N} \left(\tau_j - \sum_{k=1}^{12} \omega_k ProTeach_j^k \right)^2 \\ \text{st} & 0 \leq \omega_k \end{array} \tag{5}$$

In (5), we weight by the inverse of the variance of the teacher effects estimates. Once we estimate the best predictor of teacher value-added, we normalize the weights to sum to 1 (Mihaly et al., 2013b).

A.4 Estimating the Misclassification Rates

In order to assess the likelihood of misclassifying teachers based on the results of the ProComp assessment, we simulate the probability that teachers who pass or fail the ProTeach assessment are in particular segments of the teacher value-added distribution. For this analysis, we calculate value-added scores for teachers with Residency Certificates who have not yet advanced to the Professional Certification. We calculate the value-added scores controlling for experience to account for any differences between teachers due purely to experience effects. We then calculate the empirical Bayes estimates as described above.

We conduct Monte Carlo simulations to estimate the probability that a teacher who passes or fails the ProTeach assessment is in each decile of the teacher effectiveness distribution. In particular, we simulate draws of the entire set of teachers from the empirical Bayes distribution of teacher effectiveness:

$$\tau_j \sim N\left(\hat{\tau}_j, \frac{\sigma_\tau^2 \sigma^2(\hat{\tau}_j)}{\sigma_\tau^2 + \sigma^2(\hat{\tau}_j)} \right).$$

That is, each teacher's simulated value-added is centered around her empirical Bayes estimate with variance equal to the variance of the empirical Bayes estimate. For each draw of the teacher effectiveness data, we compute the decile in the distribution for each teacher and

then calculate the empirical probability that a teacher who passes or fails the assessment has a value-added score in each decile. Our final estimates are the averages over 500 simulated draws of the data.

B Additional Empirical Results

Table 7: ProTeach Status and Student Achievement

	Math	Reading
Passing score on ProTeach portfolio	0.022	-0.002
	(0.016)	(0.011)
Failing score on ProTeach portfolio	-0.024	-0.052**
	(0.040)	(0.022)
Incomplete ProTeach portfolio	-0.042	-0.010
	(0.040)	(0.023)
Passing score - Failing score	0.045	0.050**
	(0.043)	(0.024)

Notes: "Passing score on ProTeach portfolio" indicates that the the teacher receives a passing score on the first ProTeach submission. "Failing score on ProTeach Portfolio" indicates that the teacher received a failing score on the first ProTeach submission. "Incomplete ProTeach Portfolio" indicates that the teacher submits at least one entry but does not complete the ProTeach Portfolio. The difference between teachers passing and failing on the first attempt is the coefficient on passing minus the coefficient on failing and is given in the last row under "Passing score - Failing score." Standard errors clustered at the teacher level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: ProTeach Assessment Performance by Classroom Characteristics

	Math		Reading	
Baseline math score	0.220	-0.255	-0.508	-0.276
	(1.281)	(1.207)	(1.646)	(1.533)
Baseline reading score	-0.453	0.046	0.653	0.223
	(1.551)	(1.517)	(1.393)	(1.355)
Female	-4.197	-4.673	-2.471	-1.656
	(3.682)	(3.662)	(3.551)	(3.862)
Amer. Indian	4.282	3.668	0.575	-0.977
	(4.773)	(4.351)	(9.326)	(9.388)
Asian	-1.824	-1.789	-0.737	-0.755
	(2.361)	(2.475)	(1.887)	(1.898)
Black	-0.526	-2.070	6.105	6.373
	(4.387)	(4.349)	(4.524)	(4.857)
Hispanic	-3.030	-3.971*	-0.194	-0.333
	(2.152)	(2.186)	(2.767)	(2.685)
Multiracial	-2.654	-4.270	2.174	3.220
	(5.013)	(5.544)	(5.563)	(6.089)
Learning disabled	3.295	3.511	5.975	4.798
	(4.181)	(4.295)	(4.373)	(4.544)
Gifted services	1.521	1.591	0.886	0.386
	(2.058)	(2.027)	(1.281)	(1.175)
LEP services	3.671	4.573	-1.221	-0.899
	(5.066)	(4.862)	(5.760)	(5.459)
SPED services	-1.664	-1.708	-6.170**	-5.581**
	(3.152)	(3.122)	(2.527)	(2.645)
Subsidized lunch	-2.669**	-2.322*	-2.216	-2.057
	(1.343)	(1.306)	(1.435)	(1.392)
Class size (/10)	-0.175	-0.203	-0.213*	-0.185
	(0.177)	(0.169)	(0.120)	(0.123)
Elementary classroom	4.604**	4.133*	4.534**	4.530**
	(2.127)	(2.289)	(1.873)	(1.725)
Class size x Elementary (/10)	-1.748**	-1.776**	-1.727**	-1.467**
	(0.778)	(0.789)	(0.767)	(0.726)
Value-added		2.889*		2.901
		(1.459)		(2.068)
Test of joint sig. (p-value)	0.034	0.024	0.014	0.022
N	113	113	98	97

Notes: Regression of ProTeach assessment scores on the mean characteristics of a teacher's classroom during the testing year. Standard errors clustered at the teacher level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Models with School Fixed Effects

	Math			Reading		
Pass	0.011			-0.006		
	(0.013)			(0.010)		
Fail	0.007			-0.051*		
	(0.041)			(0.028)		
Incomplete	-0.029			0.013		
	(0.037)			(0.028)		
Composite		0.002			0.015	
		(0.014)			(0.011)	
Entry 1			0.045			-0.003
			(0.036)			(0.027)
Entry 2			-0.012			0.014
			(0.038)			(0.031)
Entry 3			-0.023			-0.013
			(0.037)			(0.025)
Pass - Fail	0.003			0.045		
	(0.043)			(0.030)		
F-test of joint sig.			0.735			0.104
p-value			0.531			0.958
N	730877	730877	730877	687664	687664	687664

Standard errors clustered at the teacher level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 10: ProTeach Composite Score and Student Achievement in the Submission Year

	Math			Reading		
Pass	0.036			0.061		
	(0.055)			(0.047)		
Fail	-0.069			-0.029		
	(0.074)			(0.064)		
Composite		0.054**			0.023	
		(0.026)			(0.018)	
Entry 1			0.013			-0.005
			(0.027)			(0.022)
Entry 2			-0.004			0.014
			(0.024)			(0.026)
Entry 3			0.103***			-0.001
			(0.036)			(0.031)
Pass - Fail	0.104			0.090*		
	(0.067)			(0.049)		
F-test of joint sig.			2.703			0.115
p-value			0.048			0.951
N	4922	4922	4922	4148	4148	4148

Standard errors clustered at the teacher level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 11: Student Achievement with Leads of Teacher ProTeach Status

	Math	Reading
Lead: passing score on ProTeach portfolio	0.032**	0.008
	(0.013)	(0.011)
Lead: failing score on ProTeach portfolio	0.012	0.045
	(0.047)	(0.028)
Lead: incomplete ProTeach portfolio	0.054	0.015
	(0.037)	(0.032)
Lead pass - Lead fail	0.020	-0.038
	(0.047)	(0.030)

Standard errors clustered at the teacher level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 12: Highest ProTeach Outcomes and Student Achievement

	Math			Reading		
Pass	0.024			-0.002		
	(0.015)			(0.011)		
Fail	-0.052			-0.052**		
	(0.048)			(0.026)		
Incomplete	-0.065			-0.026		
	(0.041)			(0.024)		
Composite		0.018			0.017	
		(0.018)			(0.011)	
Entry 1			0.044			-0.026
			(0.042)			(0.027)
Entry 2			0.035			0.023
			(0.047)			(0.030)
Entry 3			-0.050			0.017
			(0.046)			(0.026)
Pass - Fail	0.075			0.050*		
	(0.050)			(0.028)		
F-test of joint sig.			1.832			1.190
p-value			0.139			0.312
N	730877	730877	730877	687664	687664	687664

Standard errors clustered at the teacher level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 13: Full Regression Results for Baseline Models

	Math		Reading	
	(1)	(2)	(3)	(4)
Passing score on ProTeach portfolio	0.022		-0.002	
	(0.016)		(0.011)	
Failing score on ProTeach portfolio	-0.024		-0.052**	
	(0.040)		(0.022)	
Incomplete ProTeach portfolio	-0.042		-0.010	
	(0.040)		(0.023)	
Complete ProTeach portfolio		0.017		-0.010
		(0.015)		(0.010)
Standardized composite score		0.007		0.018*
		(0.018)		(0.011)
Math pretest	0.669***	0.669***	0.289***	0.289***
	(0.002)	(0.002)	(0.002)	(0.002)
Math pretest sq.	0.003***	0.003***	-0.010***	-0.010***
	(0.001)	(0.001)	(0.001)	(0.001)
Math pretest cu.	-0.021***	-0.021***	-0.007***	-0.007***
	(0.000)	(0.000)	(0.000)	(0.000)
Read pretest	0.176***	0.176***	0.499***	0.499***
	(0.001)	(0.001)	(0.002)	(0.002)
Read pretest sq.	0.001**	0.001**	-0.046***	-0.046***
	(0.001)	(0.001)	(0.001)	(0.001)
Read pretest cu.	-0.006***	-0.006***	-0.016***	-0.016***
	(0.000)	(0.000)	(0.000)	(0.000)
Female student	-0.039***	-0.039***	0.125***	0.125***
	(0.001)	(0.001)	(0.002)	(0.002)
Amer. Indian student	-0.058***	-0.058***	-0.070***	-0.070***
	(0.005)	(0.005)	(0.006)	(0.006)
Asian student	0.104***	0.104***	0.020***	0.020***
	(0.003)	(0.003)	(0.003)	(0.003)
Black student	-0.081***	-0.081***	-0.023***	-0.023***

Continued on next page

Table13 – continued from previous page

	Math		Reading	
	(1)	(2)	(3)	(4)
	(0.003)	(0.003)	(0.004)	(0.004)
Hispanic student	-0.034***	-0.034***	-0.026***	-0.026***
	(0.002)	(0.002)	(0.002)	(0.002)
Multiracial student	-0.007**	-0.007**	0.004	0.004
	(0.003)	(0.003)	(0.004)	(0.004)
Learning disabled student	-0.042***	-0.042***	-0.096***	-0.096***
	(0.004)	(0.004)	(0.005)	(0.005)
Gifted student	0.251***	0.251***	0.182***	0.182***
	(0.006)	(0.006)	(0.006)	(0.006)
LEP student	-0.062***	-0.062***	-0.158***	-0.158***
	(0.004)	(0.004)	(0.004)	(0.004)
SPED student	-0.136***	-0.136***	-0.140***	-0.140***
	(0.003)	(0.003)	(0.004)	(0.004)
FRPL student	-0.065***	-0.065***	-0.076***	-0.076***
	(0.002)	(0.002)	(0.002)	(0.002)
Class mean math pretest	0.006	0.006	-0.094***	-0.094***
	(0.009)	(0.009)	(0.007)	(0.007)
Class mean read pretest	0.077***	0.077***	0.144***	0.144***
	(0.011)	(0.011)	(0.008)	(0.008)
Class mean female	0.005	0.005	0.075***	0.075***
	(0.021)	(0.021)	(0.017)	(0.017)
Class mean Amer. Ind.	-0.092**	-0.092**	-0.119***	-0.119***
	(0.043)	(0.043)	(0.036)	(0.036)
Class mean Asian	0.228***	0.229***	0.077***	0.077***
	(0.027)	(0.027)	(0.018)	(0.018)
Class mean Black	-0.045	-0.045	-0.082***	-0.082***
	(0.031)	(0.031)	(0.023)	(0.023)
Class mean Hispanic	0.065***	0.064***	0.004	0.004
	(0.017)	(0.017)	(0.013)	(0.013)
Class mean multiracial	0.005	0.004	-0.046	-0.047

Continued on next page

Table13 – continued from previous page

	Math		Reading	
	(1)	(2)	(3)	(4)
	(0.036)	(0.036)	(0.029)	(0.029)
Class mean learning disabled	0.107***	0.106***	0.016	0.016
	(0.040)	(0.040)	(0.037)	(0.037)
Class mean gifted	-0.037**	-0.037**	-0.075***	-0.075***
	(0.017)	(0.017)	(0.013)	(0.013)
Class mean LEP	0.161***	0.161***	0.089***	0.089***
	(0.028)	(0.028)	(0.022)	(0.022)
Class mean SPED	-0.132***	-0.132***	-0.053**	-0.053**
	(0.029)	(0.029)	(0.026)	(0.026)
Class mean FRPL	-0.098***	-0.098***	-0.070***	-0.070***
	(0.014)	(0.014)	(0.010)	(0.010)
Elementary classroom	0.012	0.012	-0.017	-0.017
	(0.019)	(0.019)	(0.014)	(0.014)
Class size	-0.000	-0.000	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
Elementary x class size	0.000	0.000	0.001	0.001
	(0.001)	(0.001)	(0.000)	(0.000)
N	730877	730877	687664	687664