The Gateway to the Profession:

Assessing Teacher Preparation Programs Based on Student Achievement

Dan Goldhaber, Stephanie Liddle, & Roddy Theobald

Center for Education Data & Research, University of Washington Bothell

We gratefully acknowledge the use of confidential data from Washington State supplied by the Office of Superintendent for Public Instruction (OSPI). This research was made possible in part by a grant from the Carnegie Corporation of New York and has benefited from helpful comments from Joe Koski, John Krieg, Dale Ballou, Jim Wyckoff, Steve Rivkin, Duncan Chaplin, Margit McGuire, Jon Wakefield, and Cory Koedel. Finally, we wish to thank Jordan Chamberlain for editorial assistance. The statements made and views expressed in this paper do not necessarily reflect those of the University of Washington Bothell, Washington state, or the Carnegie Corporation. Any and all errors are solely the responsibility of the authors.

The suggested citation for this working paper is:
Goldhaber, D., Liddle, S. & Theobald, R. (2012). The Gateway to the Profession: Assessing Teacher Preparation Programs Based on Student Achievement CEDR Working Paper 2012-4. University of Washington, Seattle, WA.

Abstract: With teacher quality repeatedly cited as the most important *schooling* factor influencing student achievement, there has been increased interest in examining the efficacy of teacher training programs. This paper presents the results of research investigating the relationship between teachers who graduate from different training programs and student achievement on state reading and math tests. Using a novel methodology that allows teacher training effects to decay, we find that training institution indicators explain a statistically significant portion of the variation in student achievement in reading, but not in math. Moreover, there is evidence that graduates from some specific training programs are differentially effective at teaching reading than the average teacher trained out-of-state and that these differences are large enough to be educationally meaningful.

"Under the existing system of quality control, too many weak [teacher training] programs have achieved state approval and been granted accreditation" (Arthur Levine, former president of Teachers College, Columbia University, 2006).

"By almost any standard, many if not most of the nation's 1,450 schools, colleges and departments of education are doing a mediocre job of preparing teachers for the realities of the 21st century classroom" (Arne Duncan, U.S. Secretary of Education, 2009).

I. Teacher Training and Student Achievement

The perceived lack of quality control within the teacher preparation system paints a discouraging picture of the system's prospects for improving the teacher workforce and has led to calls for reform that include monitoring programs more closely and holding them accountable for student achievement results.[1] Evaluating teacher training programs based, at least in part, on the student performance of their trainees has already emerged as an important education reform strategy in several states such as Colorado, Louisiana, Texas, and Tennessee, and was a central tenet of the Race to the Top (RttT) grant competition.[2]

The value of teacher training is a hotly debated topic in education. Much of this debate is fueled by comparisons of teachers who hold either a traditional or alternative license.[3] Teacher training, however, is often painted with a broad brush, despite the fact that there are over 2000 traditional teacher training programs in the United States. Rhetoric about teacher training aside,

[1] See, for instance, Cochran-Smith and & Zeichner (2005), Crowe (2010), Duncan (2010), Levine (2006), NCATE (2010), Teaching Commission (2006).

[2] The U.S. Department of Education is also currently working to change regulation of teacher training programs with an aim "to reduce input-based reporting elements that are not strong indicators of program effectiveness or safety and replace them with three categories of outcome-based measures [including]…student growth of elementary and secondary school students taught by program graduates" (U.S. Department of Education 2011).

[3] See, for instance, Darling-Hammond (1999), Goldhaber and Brewer (2000), Glazerman et al. (2006). For a more thorough review of this literature, see Harris and Sass (2007).

there is relatively little quantitative information linking programs with the quality of their graduates, or research on specific approaches to teacher preparation are related to the effectiveness of teachers (National Research Council, 2010).[4] The considerable pushback against holding teacher training programs accountable for student growth-based estimates (e.g., value-added) of their graduates has focused on the shortage of research on measures of teacher training program effectiveness (Sawchuk, 2012).

Researchers have only recently begun using administrative databases to link teacher preparation programs program to in-service teachers and then to student achievement in order to draw conclusions about the efficacy of teacher training programs (Boyd et al., 2009; Henry et al., 2011; Koedel et al., 2012; Mihaly et al., 2011; Noell et al., 2008;).

Boyd et al. (2009), the only published large-scale quantitative study focused on teacher training institutions, examines training programs for teachers employed in New York City. This study suggests that there is important variation in the effectiveness of teachers graduating from different programs and, moreover, that some program characteristics (e.g., timing of student teaching) predict program effectiveness. The difference between teachers from the average institution and highest performing institution is about as large as the average difference between students who are eligible for free or reduced-price lunch and students who are not. This degree of variation is similar for both math and language arts. Furthermore, institutions that produce effective math teachers also tend to produce effective language arts teachers.

Mihaly et al. (2011) and Koedel et al. (2012) focus on the importance of empirical specification for interpreting both the point estimates and statistical significance of training program effects. Mihaly et al. focus on the implications of including fixed effects designed to

[4] As we note below, it is difficult, if not impossible to definitively assess the causal impact of training institutions on teacher candidates since the effectiveness of in-service teachers is likely to depend on both their individual attributes and what they learned while being trained.

account for (time-invariant) school context factors (e.g., an effective principal). In particular, they conclude that in order for models that employ school fixed effects to produce unbiased program estimates, data must meet two assumptions: identifiability and homogeneity.[5]

Koedel et al. (2012) further caution that studies that fail to account for the clustering of student observations for the same teacher overstate the differences between training programs because sampling variability has been inappropriately attributed to training programs.

In this paper we present research on the relationship between teacher training and teacher effectiveness (i.e., teacher-classroom-year effects) that builds on the existing literature on teacher training institutions in several ways. First, we utilize a two stage model that accounts for the clustering issue identified by Koedel et al., and test the implications of data restrictions inherent with fixed effects specifications, consistent with Mihaly et al. Second, we attempt to disentangle the contribution of training programs toward teacher effectiveness from the influence of teacher selection into particular programs.

Finally, unlike prior research, our model allows for the possibility that training program effects decay the longer a teacher is in the workforce.[6] Allowing for the possibility that training programs decay is an important feature of our model because it is quite unlikely that the impact of a teacher's training one year after the receipt of a teaching credential is the same as the impact ten or twenty years after the receipt of a credential. This is particularly important in the context of using student achievement for training program accountability purposes. One way to address the likelihood that training program effects decay with teacher experience is to only include

[5] Identifiability refers to the connectedness of training programs and schools and the representation of teachers from different preparation programs in the same schools. Homogeneity refers to the assumption that programs estimates for "highly centralized" schools (those with teachers from four or more preparation programs) are not significantly different from those for less connected schools.

[6] Henry et al. (2011) hint at the fact that program effects decay by mentioning that "the influence of colleagues, formal and informal professional development, and trial and error experiences within the classroom may significantly reduce the influence of teacher preparation that occurred 10 or 20 years earlier (p. 7)," but they do not investigate this possibility in their analysis.

novice teachers in an assessment of training programs, but this assessment will reduce the reliability of the training program estimates and almost certainly guarantee insignificant findings for smaller programs. The decay feature of our model allows more experienced teachers to contribute toward program estimates and allows for the possibility that the effect of pre-service training is not constant over a teacher's career.

We investigate training programs in Washington state and find the majority of state-accredited programs produce teachers who cannot be statistically distinguished from teachers who were credentialed outside of the state. Our estimates do, however, show some statistically significant and educationally meaningful differences in the effectiveness of teachers who received training from different programs *within* the state. The point estimates, for example, suggest that the regression-adjusted difference between teachers who received a credential from a program with the lowest performing teachers and those who received a credential from the program with the highest performing teachers is about 12 percent of a standard deviation in math and 19 percent in reading. In math, this difference is 1.5 times larger than the regression-adjusted difference in performance between students eligible for free or reduced-price lunches and those who are not; in reading the difference is 2.3 times larger.

Moreover, in the case of math instruction, but not reading, teachers employed in a school district nearby the institution where they received their teaching credential are found to be somewhat more effective than those who are employed farther away. And, for some programs, we find evidence that recent graduates from specific programs appear to be differentially effective compared to earlier cohorts from those same programs.

II. Conceptual Framework, Analytic Approach, and Data

We posit a conceptual model in which the effectiveness of teacher *j* at time *t*, τ_{jt}, is assumed to be dependent on four components: (1) individual specific (time invariant) teaching ability, φ_j; (2) the match between teachers and their schools and districts, η_{jk}; (3) a non-linear function *f* of their experience in the labor market, Exp_{jt}; and (4) the quality of the teacher-training they received, γ_{jp}. This program effect decays at a rate λ according to a decay function *g* of teacher experience:

$$\tau_{jt} = \varphi_j + \eta_{jk} + f(Exp_{jt}) + g(\lambda Exp_{jt})\gamma_{jp} \qquad (1)$$

The components of this stylized model are broadly consistent with what is observed in the teacher workforce. First, there appears to be considerable heterogeneity of teacher effectiveness that is not strongly related to observable teacher characteristics or credentials (Rivkin et al., 2005) but does persist over time (Goldhaber & Hansen, forthcoming; McCaffrey et al., 2009). Second, a newer body of evidence suggests productivity effects of good teacher-school matches (Jackson, 2010; Jackson & Bruegmann, 2009) and principal effects (Branch et al., 2012). Third, there is good evidence that teachers become more effective with additional experience, particularly early on in their careers (Clotfelter et al., 2006; Rockoff, 2004). The importance of the final component, teacher training, is certainly suggested by the research cited above (e.g., Boyd et al., 2009) but, to our knowledge, our study is the first to consider the possibility that these effects decay over time.

The objective of our analysis is to isolate the effect of training programs from the other three components in equation (1). We attempt to do this using a two-stage modeling approach. In the first stage, we estimate a standard value-added model designed to isolate the contribution of teachers toward student achievement. And in the second stage we estimate the relationship between the estimates of teacher effectiveness, derived from the first stage, and teacher training.

There is a growing body of literature that uses value-added models (VAMs) in an attempt to identify causal impacts of individual teachers on student learning, measured by standardized tests.[7] We utilize the following value-added model:

$$A_{ijst} = A_{i(t-1)}\alpha + X_{it}\beta + \tau_{jt} + \varepsilon_{ijst} \qquad (2)$$

In (2), i represents students, j represents teachers, s represents subject area (math or reading), and t represents the school year. Student achievement A_{ijst}, is regressed against: prior student achievement in math and reading, $A_{i(t-1)}$, a vector of student and family background characteristics (e.g., sex, race/ethnicity, disability, special-ed status, free or reduced-price lunch status), X_{it}, and teacher-classroom-year effects, τ_{jt}.[8]

While the model above is a commonly used methodology to derive teacher-classroom-year effects, there is no universally accepted estimation specification for this purpose (NRC, 2010), and empirically derived program estimates involve making a number of strong assumptions about the nature of student learning.[9] In particular, in (2), the teacher-classroom-year effects implicitly include any school- or district-level factors (e.g., the effectiveness of principals, the curriculum of a school district) that influence student achievement; we discuss this issue at greater length below.

[7] See Koedel et al. (2012), Aaronson et al., (2007), Boyd et al., (2009), Clotfelter et al. (2007), Goldhaber (2007), Rockoff (2004) as examples of studies that attempt to isolate the impact of schooling inputs from other factors (such as family background or class size) that influence student growth on standardized tests.

[8] We fit equation (2) using the fese command in Stata (Nichols 2008). This allows us to save the teacher fixed effects and their standard errors. Note that this command produces both cluster-robust standard errors (CRSEs) and heteroskedasticity-robust standard errors (HRSEs). Since CRSEs are not designed for getting good estimates of the sampling variablity on the fixed effects themselves, we use the HRSEs as the standard errors in the generalized least squares in equation (3) below. (We thank Mike Hansen and Austin Nichols for personal communication on this subject, November 2011.)

[9] For a discussion of this in relation to the derivation of individual teacher effects, see Todd and Wolpin (2003) and Rubin et al. (2004).

In the second stage, estimated teacher effectiveness at time t ($\hat{\tau}_{jt}$) is assumed to depend on experience level at time t (Exp_{jt}), time-invariant teacher characteristics (T_j), time-varying classroom characteristics (C_{jt}) and the training that was received while at program P:

$$\hat{\tau}_{jt} = \beta_0 + \sum_{y=0}^{4} \delta_y 1_{\{y \le Exp_{jt} \le y+1\}} + T_j\varphi + C_{jt}\kappa + e^{-\lambda(Exp_{jt})} \sum_p \gamma_p P_{jp} \qquad (3)$$

In (3), we explicitly model two of the four components in the conceptual model (1). Since prior work (Clotfelter et al., 2006; Rockoff, 2004) suggests that return to experience is non-linear and strongest early in a teacher's career, we model the return-to-experience function *f* in equation (1) by including dummy variables for each of a teacher's five years in the classroom. Each parameter of interest, γ_p, defines the return associated with receiving training at program *p* (P_{jp} = 1 if teacher *j* attended program p and 0 otherwise). A key difference between our model and models in the existing literature on training program effects is that we assume program effects decay (all training program effects decay at the same exponential rate λ) during the time that teachers are in the labor market.

The rate of decay is defined by λ, and our choice of decay function (*g* in equation (1)) is ubiquitous both in the physical sciences (e.g., Karro, et al., 2007; Wennmalm & Sanford, 2007) and economics literature (e.g., Bechtold & Sumners, 1988; Gatheral, 2010; Obizhaeva & Wang, 2005; Padmanabhan &Vrat, 1990). Note that the effectiveness of a totally novice teacher (Exp_{jt} = 0) is influenced solely by individual teaching talent and by training program, but, assuming a positive value of λ, the influence of training programs on teacher effectiveness diminishes the longer that a teacher is in the field.[10] If training program effects do not decay, then $\lambda = 0$.[11]

[10] The closest analogue we have found in the education production function literature is a value-added model proposed by Harris and Sass (2006) that allows for geometric decay in the impact of schooling inputs over time. We

Because the dependent variable is an estimate from model (2) and ordinary least squares regression may produce standard errors that are too small, we use a generalized least squares approach that accounts for the uncertainty in the dependent variable by weighting observations in proportion to the reliability of each individual estimated teacher-year effect (Aaronson et al. 2007; Borjas & Sueyoshi 1994; Koedel & Betts 2007).[12]

The challenge in interpreting the results from model (3) is that we do not control for the first two components of our conceptual model (1), pre-service teaching ability (φ_j) or teacher/school/district match (η_{jk}). These effects are likely to be confounded with training programs effects due to the non-random selection of teacher candidates into training programs and the non-random assignment of teachers into schools and districts.

In an effort to account for these sources of selection we estimate variants of model (3) that include measures of institutional selectivity, individual ability, and district or school fixed effects. These, however, are likely to be imperfect controls. Pre-service controls may be a poor proxy for individual teaching ability and it is still not totally clear that fixed effects models will yield unbiased estimates of *mean* program effects. The reason is that in a fixed-effects model, the estimates are based solely on within district or school differences in teacher effectiveness, and some of the differences between programs may be related to systematic sorting across different

---

use exponential decay because the model fit is marginally better than with geometric decay, but the correlation between the program estimates using exponential and geometric decay was over 0.99 in every specification.

[11] This model implicitly assumes that decay is a function of experience in the labor market as opposed to the time since a teacher received her training, however, we could easily replace Exp_{jt} in the exponential term with a measure of time since training. The correlation between program effects in models that use experience and models that use time since training is greater than .99.

[12] The correlations between these model results and those from unweighted, or OLS, models are all above 0.97 in both subjects. Previous work in Florida (Mihaly et al. 2011) estimates program effects with the felsdvregdm command (Mihaly et al 2010), which allows them to estimate program effects relative to the average program in the state. However, this command does not allow for the non-linear specification in model (3), so we include program dummies and estimate program effects relative to out-of-state teachers. Nonetheless, we find that the correlation between the program estimates from our model when $\lambda = 0$ (no decay) and the program estimates using felsdvregdm is over 0.9 in both math and reading.

types of districts or schools. Imagine, for instance, that there are large differences between programs, but schools tend to employ teachers of a similar effectiveness level. In this case, a school that employs teachers that are average in effectiveness, from multiple programs, would tend to have some of the least effective teachers from the best training programs and most effective teachers from the worst training programs, and thus the within-school comparison would tend to show little difference between the programs. In other words, some of the true differences between programs help explain the sorting of teachers across schools so the within-school comparisons wash out the program estimates.[13] Therefore, while the school and fixed effects models improve upon model (3) by separating the program effect from school- or district-level confounders, they introduce a comparison-group issue that makes interpretation of the results difficult.[14]

We use multiple years of data to improve the stability of the estimates, but there are two issues that arise from our use of multiple years of data. First, model (3) assumes that programs do not change over time. This may be plausible over the short term given well-documented resistance to change within academic departments (e.g., Perlmutter, 2005; Summers, 2012), but it is unlikely that the no change assumption would hold over the decades that cover the time span for which we have teachers in our sample. Thus, we also estimate additional models that include interactions between program graduate cohorts and programs to explore this issue. Second, we do not account for the possible non-random attrition of teachers from different programs. If

[13] The analogous issue arises when estimating individual teacher effectiveness and making decisions of whether or not to include school level fixed effects. We thank Jim Wyckoff for his insights on this matter. (Personal Communication, August 2011).

[14] We also attempt to account for the non-random distribution of programs *within* schools by estimating a school-level model (Hanushek et al., 1996) that regresses the average achievement of students in a school on school characteristics (e.g., enrollment, percent of students in each gender and racial/ethnic category, percent of students eligible for free or reduced-price lunches, average class size) and the percent of teachers in that school who come from each training program. The results of this specification should be robust to any non-random sorting of teachers within schools.

teachers from different programs attrite from the workforce at different rates, this will bias our program estimates, since these estimates are pooled over years. We do not address this issue explicitly in the model, but we explore the potential that non-random attrition might lead to biased program effects in the data section below.

III. Data

The data for this study is derived primarily from five administrative databases prepared by Washington State's Office of Superintendent of Public Instruction (OSPI): the *Washington State S-275* personnel report, the *Washington State Credentials* database, the *Core Student Record System (*CSRS*),* the *Comprehensive Education Data and Research System* (CEDARS), and the *Washington Assessment of Student Learning (WASL)* database.

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, all of which are used as control variables in model (3). (See **Table A1** in the appendix for means of selected teacher characteristics by training program).

The *Washington State Credentials* database contains information on the licensure/certification status of all teachers in Washington, including when and where teachers obtained their initial teaching certificates.[15,16] This database also includes teachers' test scores on

[15] From this database, we identify the institution from which a teacher received his or her first teaching certificate, which may or may not be where a teacher did his or her undergraduate work. OSPI's coding schema for first-issue teaching certificates (i.e., what we call "initial" certificates) has changed over time. Under 1961 guidelines, individuals were issued *provisional* certificates. In 1971, additional guidelines were created to issue *initial* certificates. In 2000, guidelines changed once again to the current categorization of *residency* certificates. Note however, that after a major guideline change there is still a period during which certificates may be issued under their former names. So, even in 2000, some individuals may have received certificates under previous guidelines. We code all initial certificates to account for these historical changes.

the Washington Educator Skills Test-Basic, or WEST-B, a standardized test that all teachers must pass prior to entering a teaching training program.[17]

Information on teachers in the S-275 and the Washington State Credentials database can be linked to students via the state's *CSRS*, *CEDARS*, and *WASL* databases. The CSRS 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 for the 2005-06 to 2008-09 school years. All of these variables are included as student controls in model (2). In 2009-10, CEDARS replaced the CSRS database. It contains all individual student background characteristics, but in addition, includes a direct link (a unique course ID within schools) between teachers and students. The WASL database includes achievement outcomes on the WASL, an annual state assessment of math and reading given to students in grades 3 through 8 and grade 10. (See **Table A2** in appendix for means of selected student characteristics by training program).

Like every state, Washington has requirements for initial entry into the teacher workforce, but unlike a number of states, Washington's standards are relatively stringent in the sense that it has not relied on alternative routes, such as Teach for America, as a significant source of new teachers (National Council on Teacher Quality, 2007). The great majority of the

[16] The "recommending agency" variable in these data identifies the college/university that did all of the legal paperwork to get an individual issued a teaching certificate. Thus, while it is likely that the recommending institution was also the institution where teachers were trained, the variable itself does not necessarily mean that the person graduated from the recommending agency.

[17] Since August 2002, candidates of teacher preparation programs in Washington state have been required to meet the minimum passing scores on all three subtests (reading, mathematics, and writing) of the WEST-B as a prerequisite for admission to a teacher preparation program approved by the PESB. The same is also required of out-of state teachers seeking a Washington state residency certificate. This test is designed to reflect knowledge and skills described in textbooks, the Washington Essential Academic Learning Requirements (EALRs), curriculum guides, and certification standards.

state's teachers are trained at one of the 21 state-approved programs.[18] There is, however, clearly heterogeneity in the selectivity of programs preparing teachers. For instance, the University of Washington Seattle (UW Seattle) is considered the flagship university in the state and in 2009, the 75th percentile composite SAT score of incoming UW freshman was about 1330. Nearly every other program in the state had lower 75th percentile SAT scores ranging between 1070 and 1290.[19] And, a few accredited programs do not require applicants to submit admissions test results in order to be considered for admission.[20] To account for this heterogeneity we use institution-level data from *The College Board*, which includes annual (since 1990) measures of selectivity based on the high school grades, standardized test scores, and admissions rates of incoming freshman.[21]

We combine the data from these sources described above to create a unique dataset that links teachers to their schools and, in some cases, their students in grades 3 through 6 for the 2005-06 to 2009-10 school years in both math and reading. Due to data limitations, not all students in grades 3–6 across these five school years can be linked to their teachers; until recently, the state only kept records of the names of individuals who proctored the state assessment to students, not necessarily the students' classroom teacher.[22] Across all these grades

[18] In each year from from 2006-07 to 2008-09, at least 95 percent of teachers in Washington state were certified via traditional training programs; the remainding teachers were certified through alternative training programs within PESB-accredited institutions of higher education (https://title2.ed.gov/Title2STRC/Pages/ProgramCompleters.aspx).
[19] The one exception is the University of Puget Sound whose composite 75th percentile SAT score was 1340.
[20] Heritage University has an open enrollment policy. City University and Antioch University both focus on adult learning and bachelor's degree completion suggesting less stringent entrance requirements.
[21] Year-to-year Pearson correlations for each of these selectivity measures are typically high, i.e., above 0.90.
[22] The proctor of the state assessment was used as the teacher-student link for the data used for analysis for the 2005-06 to 2008-09 school years. The assessment proctor is not intended to and does not necessarily identify the subject-matter teacher of a student. The "proctor name" might be another classroom teacher, teacher specialist, or administrator. We take additional measures to reduce the possibility of inaccurate matches by limiting our analyses to elementary school data where most students have only one primary teacher and only including matches where the listed proctor is reported (in the S-275) as being a certified teacher in the student's school and, further, where he or she is listed as 1.0 FTE in that school, as opposed to having appointments across various schools. And for the 2009-10 school year, we are able to check the accuracy of these proctor matches using the state's new Comprehensive Education Data and Research System (CEDARS) that matches students to teachers through a unique course ID. Our

and years we were able to match about 70 percent of students to their teachers and estimate value-added models (VAMs) of teacher effectiveness.[23]

Our analytic sample includes 8,718 teachers (17,715 teacher-years) for whom we can estimate VAMs and for whom we know their initial teacher training program as being either from one of 20 state accredited teacher preparation programs, or from outside of the state.[24] These teachers are linked to 291,422 students (388,670 student-years) who have valid WASL scores in both reading and math for at least two consecutive years. **Table 1** reports selected student characteristics for the 2008-09 school year for an unrestricted sample of students, i.e., those in grades 4–6 who have a valid WASL math or reading score but could not be matched to a teacher, and our restricted analytic sample described above. T-tests show that while nearly all of these differences are statistically significant, none of them are very large.[25]

Before moving to the main results in the following section, we note that we investigate the potential that non-random attrition from the teacher workforce might lead to biased program effects in two ways. First, we compare time-invariant teacher characteristics by training program, for new teachers in 2006-07 to the characteristics of those teachers who remained in the workforce until 2009-10. Approximately 5 percent of the within-program differences by sex, race/ethnicity, and licensure score are statistically significant at the .05-level, suggesting that the

---

proctor match agrees with the student's teacher in the CEDARS system for about 95 percent of students in both math and reading.

[23] For the 2005-06 to 2008-09 school years where the proctor name was used as the student-teacher link, student-to-teacher match rates vary by year and grade with higher match rates in earlier years and lower grades. For a breakdown of match rates by subject, grade, and year see **Table A3** in the appendix**.**

[24] Consistent with Constantine, et al. (2009), we define teacher preparation programs as those from which new teachers must complete all their certification requirements before beginning to teach. Lesley University produced its first teachers in 2009-10. So, although it is an accredited institution in Washington state, our observation window precludes it from being included in our analysis.

[25] Comparisons for other years reveal similar results.

within-program attrition is approximately random.[26] Second, we compare the average teacher effectiveness of new teachers in 2006-07, by program, to the average teacher effectiveness of those teachers who remain in our sample until 2009-10. We find few significant differences by program. But we also caution readers that these results are based on small sample sizes. For example, only two programs placed more than 30 teachers into grade 4–6 classrooms in Washington state during 2006-07, and only one of those programs retained just over 50 percent of their teachers until 2009-10. Nonetheless, results from both of these tests suggest we ought not be overly concerned that non-random attrition in our sample will bias our program effects.[27]

IV. Results

Prior to describing our findings on teacher training programs, a few peripheral findings warrant brief notice. **Table 2** shows the coefficient estimates from the first stage models used to generate the teacher-classroom-year effects. As is typically found in the literature, there are significant differences between student subgroups in achievement. And while not reported, we also estimated models that allow for non-linear relationships between current achievement and base-year test scores.[28] The teacher effectiveness estimates from these non-linear models are highly correlated (above 0.9 in both subjects) with those generated by the linear specifications reported in **Table 2**.[29]

A. Allowing for Decay of Program Effects and Testing for Selection into Program

[26] Importantly, there is no evidence that teachers with higher or lower licensure scores are disproportionately leaving the profession within the years of our study.

[27] Whether comparing observable teacher characteristics or teachers' value-added estimates over time, t-tests for samples that include all teachers are largely consistent with those including only new teachers.

[28] These models include a squared term for prior year's student achievement and results are available from the authors upon request.

[29] Since we are estimating teacher-classroom-year effects, we cannot estimate models that include class-level aggregates. But, when we estimate models using multiple years of data to inform the estimates of teacher effectiveness, thus allowing for classroom aggregates, we find that the correlation of teacher effectiveness estimates between these and the models without class-level aggregates is 0.65 in math and 0.61 in reading.

In **Table 3** we present training program effect estimates from three model specifications for math and reading: a no decay specification (columns 1 and 4 for math and reading respectively); a specification that allows for program decay (columns 2 and 5); and a specification that allows for decay and adds a number of institution-level selectivity controls to the model (columns 3 and 6). The reference category for the program effects is out-of-state teachers, meaning the program effects are interpreted as the effect of a particular program (in standard deviation units of teacher effectiveness) relative to the average effectiveness of out-of-state teachers.[30]

For our linear specification (columns 1 and 4), ANOVA models suggest that only a small proportion (less than 1 percent) of the total variation in teacher effectiveness is explained by the training program indicators, and F-tests only confirm their joint significance in reading.[31] However, these program indicators still explain roughly the same proportion of the variation in teacher effectiveness in both subjects as teacher credentials (degree, experience level) that are currently used for high stakes personnel decisions.[32]

---

[30] The specification of decay we utilize assumes that, holding teacher and classroom covariates constant, a teacher's effectiveness over time decays to the average effectiveness of out-of-state teachers *of the same experience level.* Out-of-state teachers are a sensible reference category because there are more out-of-state teachers in our sample than teachers from any individual program, and the effectiveness of out-of-state teachers is statistically indistinguishable from in-state teachers both in math (p=.355) and reading (p=.997).

[31] The p-values for joint F-tests of all program dummies in our no decay models are less than 0.01 in reading and 0.34 in math. Only 2 program indicators in math, and 5 in reading are statistically significant at the 90 percent confidence level in these models where the standard errors are clustered at the teacher level. Importantly, no more programs are statistically significant in math than we would expect due to random chance alone. This could be a sample size concern, as the majority of programs have fewer than 200 teachers in our sample, while a rough power calculation suggests that at least 600 teachers are necessary to achieve 80 percent power for the average program given the estimated magnitude and variability of the effects. Furthermore, consistent with Koedel et al, (2012), when the standard errors are clustered at the program level (as in Boyd et al., 2009), the number of statistically significant program effects increases to 17 in math and 15 in reading.

[32] We estimate both Type I and Type III sum of squares for the program indicators as this provides an upper and lower bound of the proportion of the variation in teacher effectiveness (in math and reading) explained by the training programs. In practice it matters little whether we use the Type I (0.27 percent explained in math and 0.56 percent in reading) or Type III estimates (0.24 in math and 0.40 in reading) as the difference between them is small. In estimating the Type I sum of squares, we include the program dummies before teacher credentials (but after exogenous teacher characteristics, like race and gender) so this estimate is the upper bound on the proportion of variation explained by programs.

The standard deviation of the Empirical Bayes (EB) adjusted estimates of the program indicators are about 0.01 in math and 0.02 in reading in all of the models in **Table 3**.[33] To put this in context, in reading, these differences are one quarter the size of the regression-adjusted difference between students who are and are not eligible for free or reduced-price lunches. Moreover, from model (2), the standard deviation of teacher effects is 0.20 in math and 0.16 in reading. So, the standard deviation of the program estimates is about 5 to 12.5 percent of the standard deviation of the teacher effects. This finding is far more modest than Boyd et al. (2009) who find that "a one standard deviation change in the effectiveness of the preparation program corresponds to more than a third of a standard deviation in the teacher effect for new teachers" (p. 429). One reason we find less heterogeneity in our program estimates is that Boyd et al. cluster standard errors at the program level while we follow Koedel et al. (2012) and cluster standard errors at the teacher level. We therefore obtain relatively larger standard errors, resulting in more shrinkage of the program estimates and less heterogeneity. Beyond this, we can only speculate as to why we find less heterogeneity in our program estimates than Boyd et al. It is possible these differences result from the fact that New York has more training programs (i.e., 30) than does Washington, that New York training programs draw potential teachers from a different distribution, or that training programs in Washington are more tightly regulated and are therefore more similar to one another.

There is little change in the number of significant program coefficients moving from the no decay (columns 1 and 4 for math and reading respectively) to decay (columns 2 and 5) specifications, but there are some marked changes in their magnitudes. For instance, the average

[33] To do this we calculate estimates for the program effects that use as many years of data that are available to inform each individual estimate and then adjust for variance inflation by shrinking the estimates back to the grand mean of the population in proportion to the error of the individual estimate. For more detail on this Empirical Bayes methodology, see Aaronson et al. (2007).

absolute values of the coefficient estimates in math and reading in the no decay specifications are 0.016, and 0.020 respectively, whereas the corresponding average in the decay specifications are 0.027 and 0.030. The change in magnitude is to be expected given that the interpretation of the training program coefficients is different in the decay specification. Specifically, the program indicators in the no decay model are the estimated effects on student achievement of having a teacher who received a credential from a specific training program regardless of when he/she received the credential, whereas the program coefficients in the decay models are the estimated effects for *first-year* teachers who get their teaching credentials from different programs.

In both math and reading the magnitude of the estimated decay parameter is about .05, suggesting that training program effects do decay as teachers gain workforce experience. This has important implications for how we think about the influence of training programs on teacher quality, as is illustrated by **Figure 1**, which shows how the effects of training programs for teachers with varying workforce experience decay over time base on estimates of the decay parameter (λ) from the both our decay and selectivity decay models. These results clearly suggest that teacher training should not be thought of as invariant to a teacher's workforce experience as, for instance, the "half-life" of teacher training effects is estimated to be between 12.9 and 13.7 years in math and between 11.3 to 15.5 years in reading depending on model specification.[34]

Focusing on the individual program estimates in the decay models, we see that our program rankings are broadly consistent with institutional SAT rankings in that they share modest positive correlations.[35] That said, there are a few surprises. For instance, there are relatively selective institutions (based on SAT scores) that do not appear to be graduating

[34] The estimate of the "half-life" is $\frac{\ln(2)}{\lambda}$.

[35] Correlations between program rankings and SAT scores are 0.4 in math and 0.2 in reading.

particularly effective teachers and less selective institutions that appear to be producing very effective teachers. For example, while not significant at the 95 percent level, both UW Tacoma and UW Bothell, have composite 75th percentile SAT scores of 1120 and 1150, respectively—the fourth and fifth lowest of all other institutions requiring SATs in 2009—and have graduated some of the most effective math and reading teachers in the state. Whereas more selective institutions, such as Gonzaga University, with a composite 75th percentile SAT score of 1270—ranking fourth highest—have graduated teachers who are significantly less effective in reading compared to out-of-state teachers or graduates from the majority of other in-state programs.[36]

We attempt to account for selection into training programs in our models with the inclusion of various measures of institutional selectivity: the composite (math and verbal) 75th percentile score on the SAT for incoming freshman, the percent of incoming freshman whose high school grade point average was above 3.0, and admissions rates for incoming freshman (i.e., total admitted/total applied).[37] Columns 3 and 6 in **Table 3** provide the individual training program estimates from these selectivity models. There are a handful of notable changes in individual program estimates moving from our decay models (columns 2 and 5) to those with institutional selectivity controls (columns 3 and 6). In math, for example, the coefficient for Eastern Washington University turns negative and increases in magnitude. The coefficients for St. Martin's University and Evergreen State College become increasingly negative and become statistically significant. And in reading, the coefficients for Seattle Pacific University and Western Washington University turn negative with smaller magnitudes, the latter losing

[36] Implicit in our estimates is the notion that the training programs themselves are not changing over time. We explore this issue in more detail in Section D and find some evidence of change for some training programs, but it is somewhat speculative.

[37] Since most training programs are part of four-year institutions, we subtract four years from a teacher's certification date to approximate their entry year into their certifying institution. Teachers who were certified out of state (22 percent), entered school before 1990 (31 percent), or graduated from institutions that don't report selectivity data (9 percent) are missing institutional selectivity data, but are included in the regression with a dummy indicator for missingness.

significance. However, overall there is little evidence that the inclusion of selectivity controls has an impact on the estimates. Indeed, the Pearson correlation coefficients for the program estimates of the decay model and the program estimates from the selectivity model are 0.95 for math and 0.97 for reading.

There are several potential explanations for the finding that our selectivity results differ little from the decay results. First, it may be that selection into training programs is not, at least for the workforce as a whole, a significant mechanism influencing differences between teacher training programs. Second, our controls are for the selectivity of the *institutions*, not training programs themselves, and, regardless, may be a poor proxy for individuals' pre-service capacity to teach. In fact, F-tests show that together these three selectivity measures only significantly improve model fit for reading (at the 0.10 level). Moreover, the coefficient estimates from the selectivity variables themselves do not provide a consistently strong argument that program selectivity is an important factor in predicting teacher effectiveness.[38] We attempt to better account for individual pre-service capacity/selection into program by adding an additional control for selection: a teacher's average score on the three subtests of the WEST-B.[39] The WEST-B coefficients are positive for both subjects and statistically significant in math

[38] We find a small positive effect for the composite 75th percentile SAT score of the institution in math, but a small negative effect in reading; neither effect is significant. The percent of incoming freshman whose high school GPA was above 3.0 is negative in math, but positive in reading, and again neither effect is significant at conventional levels (although the effect for reading is marginally significant). The effect of admissions rates is statistically significant at conventional levels for reading (ß = 0.001) and marginally significant for math (ß = 0.002). However, its *positive* direction is somewhat surprising given that we might expect *more* selective institutions—i.e., those with *lower* admissions rates—would graduate *more* effective teachers. We explore the admissions finding in greater detail by quantifying admissions rates as quartiles and re-running the model with the highest quartile as the referent category. Each of the coefficients for lower three quartiles are increasingly negative reiterating our rather unusual finding that teacher effectiveness is positively correlated with admissions rates.

[39] Since individuals are allowed to take the exam more than once and scores may be contingent on the number of times the test is taken, we average teachers' first scores on the three WEST-B subtests.

(marginally in reading) indicating that higher-scoring teachers tend to be more effective teachers.[40] However, the inclusion of this test does little to change the program estimates.[41]

A final possibility is that the program estimates reported in **Table 3** are biased by the sorting of teachers and students into classrooms (Rothstein, 2010). For instance, were it the case that teachers from particularly effective training institutions were systematically more likely to be teaching difficult to educate students (in ways that are not accounted for by the control variables reported in **Table 1**) and vice versa, we might expect an attenuation of the program coefficients. This possibility is explored in the next sub-section.

B. Robustness Checks: Accounting for Non-Random Sorting of Teachers and Students

We assess the potential bias associated with the non-random sorting of teachers and students by estimating models that include district and school fixed effects. These results are reported in columns 2 and 3 (for math) and columns 5 and 6 (for reading) in **Table 4** next to our decay model results copied from **Table 3** (in columns 1 and 4). Though not reported here, we also estimate a model that aggregates program and teacher effectiveness information to the school building level.[42] The argument for the fixed effects specifications is that they account for time-invariant district or school factors such as curriculum or the effectiveness of principals. The

[40] The WEST-B coefficients are 0.019 for math and 0.011 for reading.

[41] Of our full sample of 8,718 teachers, only 1,469 (16.8 percent) have WEST-B scores largely because the majority of teachers (roughly 85 percent percent) were enrolled in a program before the WEST-B requirement was put into effect. The correlation between models with and without the WEST-B test, for the subsample of teachers that have WEST-B scores, is over 0.99 for each subject. See **Table A4** in the appendix for the number of teachers with WEST-B scores who graduated from each program.

[42] The school-aggregated models regress average teacher effectivness (equally weighted by teachers) in a school against the share of teachers who hold a credential from each training program plus the other covariates in the model aggregated to the building level.

argument for the models with building level aggregation is that they wash out the potential for within-school sorting of teachers to classrooms.[43]

As described in Mihaly et al. (2011), there are two related data issues that arise when estimating fixed effects specifications of equation (3): whether schools and programs are "connected" enough in a sample such that estimates of program effects can be based on within school (or district) variation in teacher effectiveness (identifiability); and whether "highly centralized" schools (those with teachers from four or more preparation programs) are significantly different from less centralized schools such that the findings may not generalize to the full sample of schools (homogeneity). We follow Mihaly et al.'s (2011) procedures for testing for identifiability and homogeneity and conclude there is minimal concern about either in our data (at least relative to the situation they investigate in Florida). Identifiability does not appear to be a significant issue here since only about 15 percent of the schools in our sample (which employed about 5 percent of all teachers) were staffed with teachers from a single training program.[44] We cannot rule out homogeneity issues as t-tests comparing the characteristics of students and teachers in schools with teachers from one training program to those in schools with teachers from four or more training programs reveal that there are a few significant differences, but these differences are relatively small.[45]

[43] The fixed effects specifications allow for program effects to decay, but the school aggregate models do not permit this as decay is specified by the experience level of individual teachers. The correlations between estimates from these models and our no decay models are 0.63 in both subjects.

[44] This contrasts sharply with Mihaly et al's sample of Florida teachers wherein 54.1 percent of schools had teachers from a single training institution suggesting that programs in Florida serve much more localized markets than those in Washington state.

[45] Students in schools with teachers from only one training program are significantly more likely to be American Indian or white and have teachers with at least a master's degree, whereas students in schools with teachers from four or more training programs are significantly more likely to meet state math standards and be Black, Asian, and/or bilingual. Teachers in schools with only one training program are marginally (at the 10 percent significance level) more likely to have higher WEST-B scores. The t-tests are available from the authors upon request.

As is readily apparent from scanning the results across columns in **Table 4**, most of the program coefficients from our fixed effects regressions are far smaller than those reported in our decay models and are less likely to be statistically significant. For instance, the average absolute values of the program indicators in the decay models are 0.03 for both math and reading. These values attenuate to 0.01 for math and 0.02 for reading in both the district and school fixed effect models, and f-tests of the program indicators are not jointly significant in either math or reading in the fixed effect models. This finding may simply reflect the fact that there is little true difference between teachers associated with the program from which they received their teaching credential after accounting for selection into district, school, or classroom. But there are also at least two good reasons to doubt that these are the right specifications. In particular, in the case of the fixed effects models the decay parameter is estimated to be negative and marginally significant in math, and insignificant in the case of reading. This finding is seemingly implausible as it implies, in the case of math, that the effects of training programs increase with teacher experience in the workforce.

Moreover, while f-tests of the school and district fixed effects suggest that they improve the fit of the model, other measures do not. Specifically, the BIC—which penalizes for each new parameter added to the model—is lower (and thus better) for models *without* fixed effects than models with fixed effects in both subjects.[46]

Since results from models reported in Tables 3 and 4 have different interpretations, we now seek to provide guidelines for interpreting these program effects. While we believe that it is important to incorporate the decay of program effects into our model—and we see that estimation of these decay parameters improves the fit of our model—readers interested in

[46] Specifically, $BIC = -2\ln(L) + k\ln(n)$, where L is the likelihood of the model, k is the number of free parameters, and n is the sample size.

comparing the magnitude of program effects in Washington to those in New York (Boyd et al., 2009) or Louisiana (Noell et al., 2009) should use the results from the base model (columns 1 and 6 of Table 3), as these other papers do not incorporate decay. Readers interested in using programs to *predict* the effectiveness of teachers from different programs should use the estimates from the base decay model (columns 2 and 5 of Table 3), as these models combine both a program's contribution to teaching effectiveness and the latent teaching ability of teachers from that program in the program estimate. And readers interested in identifying program *contributions* to teacher effectiveness should use the estimates from the selectivity decay model (columns 3 and 6 of Table 3), as these models control for at least some of the confounding influence of the non-random selection of teachers into programs. All this said, the correlations between all these models are quite high—over 0.92 in both math and reading and therefore, these distinctions have minimal impact.

A much more difficult decision is between these three models and the district (columns 2 and 5 of Table 4) and school (columns 3 and 6 of Table 4) fixed effect specifications. The models that do not include fixed effects do not control for the shared experiences of teachers in districts or schools that are not correlated with covariates in model (3). But, the non-random sorting of teachers from programs into schools and districts complicates the interpretation of the school and district fixed effect models. Moreover, as noted above, while an F-test confirms that the addition of district and school effects improves the likelihood of the model in both math and reading, the BIC is lower (and thus better) for models *without* fixed effects than models with fixed effects in both subjects. This suggests that the addition of fixed effects results in a less parsimonious model for estimating program effectiveness.

As we discussed in the analytic approach section, there really is no way, outside of an experiment, to know which specification is appropriate. The fixed effects models in **Table 4** may be doing a better job of accounting for the types of selection that lead to non-random (controlling for covariates) matching of teachers to students. But, on the other hand, the models in **Table 4** could lead to attenuated program effects because some of the true differences between programs are correlated with teacher selection into districts, schools, and classrooms.

Despite the individual differences in some program indicators between model specifications, it turns out that there is a fair amount of agreement across many, but not all, of the specifications. As we show in **Table 5** the Pearson correlations of the estimates from all specifications in **Table 3** and the district fixed effects specifications in **Table 4** are all over 0.60 within subject. By contrast, the correlation between the estimates from the school fixed effects models and the models without fixed effects are far smaller, particularly so in math where the correlation is not statistically significant. Lastly, it is worth noting that cross-subject, within model specification, correlations are in the neighborhood of 0.4 to 0.5, suggesting that programs producing teachers who are effective at teaching math tend to also produce teachers who are effective at teaching reading.

C. Investigating the Potential of Program Specialization

The above findings provide program estimates in general but do not allow for the possibility that training programs may specialize in preparing teachers to serve particular types of students.[47] Indeed, as Boyd et al. (2009) point out, the degree to which the quality of training programs is contingent on where a teacher works (i.e., that certain programs may serve districts with particular student populations particularly well), begs important questions about teacher

[47] The above results also do not consider the possibility that training institutions and school systems may collaborate in ways that lead to school district specialization (e.g., student teaching, curricula use).

preparation. For example, Heritage University's website notes that the institution graduates "more English Language Learner endorsed educators than any other institution in the state."[48] And UW Seattle's teacher preparation program promotes itself as offering "fieldwork opportunities in [a] network of partner schools, all located in culturally diverse urban communities around the Puget Sound area."[49]

Since the main results presented in this paper are focused at the teacher level, we do not investigate whether there are differences in program effectiveness for different subgroups of students. However, using the same data, Goldhaber & Liddle (2011) estimated a student-level model similar to our model (1) with additional teacher, classroom, school, and district covariates and interactions between program indicators and selected student characteristics (i.e., whether a student is eligible for free lunch, is receiving LEP services, is Black, is Asian, and/or is Hispanic). They find relatively little evidence of such specialization.

In addition to having differential effectiveness by student subgroups, it may be that teachers from particular programs are better prepared to teach students in or near the school districts where they were trained because pre-service teachers work extensively with local in-service teachers in the field as part of their practicum training. This would make sense since there is evidence that teacher labor markets are quite localized. Boyd et al. (2005), for example, report that "an individual is twice as likely to teach in a region that is within five miles of his or her hometown as one 20 miles away and about four times as likely to teach in a region within five miles of his or her hometown as one 40 miles away" (pg. 123). To test how proximity to training institutions teacher effectiveness in our sample, we calculate three mutually exclusive proximity indicators to capture whether a teacher, in any given year, taught in school district

[48] Source: http://www.heritage.edu/LinkClick.aspx?fileticket=sBIF4rVDnV0%3d&tabid=269
[49] Source: http://education.washington.edu/areas/tep/

within 10, 25, or 50 miles of their training institution and then re-estimate our base model including these proximity indicators.[50] Results from our base decay models that include indicators for teacher employment proximity to training institution are inconsistent across math and reading achievement. For math teachers, the linear proximity effects are positive and significant for teachers who taught within 10 or 25 miles of their training program suggesting that such teachers are more effective than those who teach further away. One might think, therefore, that training programs and local school districts somehow collaborate on curricula use or pedagogy thus making graduates from nearby programs especially effective teachers for local student populations. Thus, unlike Boyd et al. (2005), we find some evidence of geographic specialization. However, we caution that this finding for math is tempered by the results from the reading model, which shows a negative and at least marginally significant effect for teaching within 50 miles of one's training institution.[51]

D. Testing for Programmatic Change

It is quite possible that teacher training programs change over time due to state mandates or institutional initiatives designed to improve teacher training and subsequent teacher effectiveness across the board. In 2008, for example, the PESB implemented the WEST-E, a content knowledge exam required for all candidates applying for endorsements on their initial teaching certificate designed to more fully align with Washington standards than was the case with previous exams. The expectation is that teachers who were required to pass this exam

[50] These distances were calculated using ESRI's ArcGIS software. More specifically, the Generate Near Table tool in ArcMap calculates the distance between an XY point (i.e., a training institution) and the closest line of a polygon (i.e., a school district) within a specified search radius. If any portion of a school district falls within the given radius, it is included in that proximity measure.

[51] It could be that we are unable to detect geographic proximity effects due to a lack of geographic specificity in the certification data. Although we know the name of the institution that issued a teacher's credential, we do no know the specific, physical site a teacher attended to obtain that credential. Several institutions have multiple sites (e.g., the number of total sites by program are: Central University: 6, City University: 7, Eastern Washington University: 2, Heritage University: 5, Saint Martin's University: 3, Washington State University: 4, and Western Washington University: 4.)

should be more effective in the classroom (especially in math) than teachers from earlier cohorts who were not held to such a standard. We cannot currently test whether this requirement has affected the quality of the teacher workforce since the test was only required of teachers who could have begun teaching in 2009-10 and too few of these teachers are in our analytic sample, but we will be able to assess this as new cohorts of teacher data become available.

It is also the case that the relative magnitude of individual program estimates could change over time due to institutional changes of individual programs.[52] To test this, we estimate variants of our base model that include interaction terms between program dummies and grouped cohorts of recent graduates.[53] More specifically, we focus on two grouped cohorts: (1) graduates between 2005 and 2009 and (2) graduates between 2000 and 2004. There is nothing special about those particular years. However, we wished to pick a time span over which the program estimates would be informed by a reasonably large number of teachers (at least 30).[54]

Prior to reporting whether individual program estimates have changed over time, it is worth noting that we compare the effectiveness of teachers trained in Washington state, as a whole, to those who received their training out-of-state and credentials from OSPI, and, in particular, whether there was evidence of changes in relative effectiveness over time. To do this, we estimate a version of our linear base model which substitutes the individual program indicators for a single in-state training indicator variable and also includes interactions between

[52] In 2003, for example, the University of Washington received a matching grant for 5 million dollars over a five-year period (2003 to 2008) from The Carnegie Corporation of New York's Teachers for a New Era Project to implement and examine changes in its teacher training program. However, future teachers from UW did not immediately experience such changes. For example, the four cohorts from 2004-05 to 2007-08 each received only parts of the renewed program. The first cohort in the fully renewed program got their certification in spring 2009. Because of the induction year, they didn't finish the full program until spring 2010.

[53] Interaction effects from models with district or school fixed effects look similar to those from the base models and all have Spearman rank correlations above 0.87 in math and 0.72 in reading.

[54] Because some programs graduated few (if any) graduates during these years, we focus on the twelve programs that graduated at least 30 new teachers during the years specified by each cohort grouping. All other programs are collapsed and their combined effects are reported.

this variable and having been trained recently, i.e., whether a teacher was certified within the last five years (2005 to 2009) or the five years prior to that (2000 to 2004).[55] The interaction terms for both subjects and both certification cohorts are positive, however, only the interaction for the 2005-09 cohort in reading is statistically significant.[56] Nonetheless, these results suggest that teachers who were trained in Washington state within the last five to ten years may be relatively more effective than those who had been credentialed in-state prior to 2000, at least as compared to teachers who were credentialed by OSPI.

We now turn to changes in individual program effectiveness. **Table 6** shows results for two linear models (math and reading) that include interaction effects for each of the programs by cohort grouping.[57] Columns 1 and 4 give the main program estimates for graduates before 2000 in math and reading respectively. Columns 2 and 5 show the interaction effects for teachers who graduated from each program between 2000 and 2004. Columns 3 and 6 show the interaction effects for teachers who graduated from each program between 2005 and 2009.[58] As we might expect given the above finding regarding the effectiveness of teachers who are trained in-state versus out-of-state, some of the interaction effects are positive and significant for each subject and cohort grouping. This suggests that, for some programs, more recent cohorts of teachers who graduated from a particular program are more effective than teachers who graduated from

[55] F-tests show that these interaction terms are jointly significant for both math and reading (p=0.02), but not for math (p=0.35) models.
[56] For math, the coefficient for the interaction between being trained in Washington and being certified between 2005 and 2009 is 0.021 with a standard error of 0.015. The coefficient for the interaction with the 2000 to 2004 cohort is 0.010 with a standard error of 0.014. For reading, the interaction effect for the in-state trained 2005-09 cohort is 0.037 with a standard error of 0.013. This same interaction term for the 2000-04 cohort is 0.008 with a standard error of 0.013.
[57] F-tests show that these interaction effects are jointly significant for both math and reading (p=0.05), but not for math (p=0.41 model)s.
[58] The total estimated effectiveness (relative to a teacher who is trained out-of-state and, hence, whose credential is obtained from OSPI) for a teacher in a recent cohort is the *sum* of the main effective and the cohort-time period interaction term. For instance, relative to a teacher trained out-of-state, the estimated effectiveness for math teachers who obtained a credential from Seattle University in the 2005-09 time period is .05, the sum of the main effect (-0.05) and the 2005-09 interaction term (0.10).

that program before 2000 relative to changes in effectiveness for teachers trained outside of Washington. For example, reading teachers trained at Seattle University and Whitworth University before 2000 were significantly *less* effective than teachers trained out-of-state during that same time period. By 2000-2004, graduates from each university were significantly *more* effective at teaching reading compared to prior graduates from those programs (relative to any changes in effectiveness for out-of-state teachers) and have continued to maintain this level of relative effectiveness into 2005-2009.[59]

Thus, we find some evidence that the relative effectiveness of graduates from various programs has changed over time and, as a whole, teachers recently trained in-state appear relatively more effective than those who received training in the past in terms of reading instruction. However, all programs are measured relative to those teachers who received training out-of-state, so we cannot say whether this is a reflection of the effectiveness of teachers who received an in-state credential or the possibility that there is a change in the quality of teachers who are coming in from out-of-state.[60] Further, we cannot distinguish between changes in programs and changes in the quality of graduates from these programs.

V. Discussion and Conclusions

U.S. teacher policy is on the cusp of significant changes. Much of this focuses on in-service teachers, and the linking of student growth measures to the evaluation of teachers. But there is increasing interest in relating student outcomes to teacher training with the hope that identified differences in program effects ultimately lead to more effective selection and training

[59] Eastern Washington University and UW Bothell follow a similar trend to Whitworth and Seattle Universities. We find little evidence that graduates from programs become less effective over time relative to earlier cohorts from their same programs. Teachers trained at Antioch University before 2000, for example, were just as effective as out-of-state teachers in the same time period for both math and reading. By 2000-2004, math teachers appear to be less effective than previous graduates. However, this difference does not persist to the 2005-2009 cohort where teachers from Antioch are once again statistically indistingishable from out-of-state teachers trained before 2000.

[60] And note that even for recent cohorts of teachers, out-of-state trained teachers are still relatively effective compared to those who received an in-state credential.

practices for pre-service teachers. The results presented in this paper represent one of the first statewide efforts to connect individual teacher training institutions to student achievement and explore the extent to which model specification influences estimated results.[61]

In general our findings suggest that where teachers are credentialed explains only a small portion of the overall variation in the effectiveness of in-service teachers. This is now a common finding in the educational productivity literature; it appears that the best assessments of teachers are those based on actual classroom performance rather than pre- or in-service credentials. That said, the differential in the effectiveness of the teachers credentialed by various programs is meaningful. For instance, the regression-adjusted difference between teachers who receive a credential from the least and most effective programs is estimated to be 3.9 to 13.4 percent of a standard deviation in math and 9.2 to 22 percent of a standard deviation in reading.[62] To put this in context, in decay models with no fixed effects, the average expected difference in student performance between having a math teacher from the most effective program and the least effective program is at least 1.5 times the regression-adjusted difference between students who are eligible for free or reduced-price lunches and those who are not ($\beta_{FRPL} = -0.076$). For reading, this difference is over two times the difference between students with learning

[61] It is important to note several caveats about the analyses we presented here. First, the samples used to detect differential program estimates for student subgroups and programmatic change over time were relatively small so it is conceivable that effects do exist but their magnitude is too small to detect with the sample at hand. Second, our analyses are focused entirely on elementary schools and teachers. It is conceivable that comparable analyses would yield results that look quite different at the secondary level. Third, students' outcomes on standardized tests are only one measure of student learning, so it is possible that value-added approaches miss key aspects of what different training institutions contribute to teacher candidates. Finally, our decay models assume that a teacher's performance decays to the average performance of out-of-state teachers of the same experience level. What might be more plausible, however, is that teacher performance decays to the average performance of teachers *at the same school* (teacher "acculturation"). Unfortunately, our specification of decay requires program effects to be measured relative to a reference category, so we cannot modify our model to allow teacher effects to decay to an average teacher, either across the workforce or within the same school. In the future it will be possible to estimate more flexible specifications that allow for differential types of decay, but this is currently infeasible given that we only have four years of value-added data for teachers in our sample.

[62] These differences across models for math and reading, respectively, are: 7.3 and 12.7 (no decay), 11.5 and 19.2 (decay), 13.4 and 22.0 (selectivity decay), 3.8 and 12.1 (district FE decay), and 4.4 and 9.2 (school FE decay).

disabilities and those without ($\beta_{disability} = -0.083$). Moreover, this same difference is larger than the size of the estimated difference between a first-year teacher and a teacher with five or more years of experience by one and a half times in math ($\gamma_0 = -0.074$) and more than 2 times in reading ($\gamma_0 = -0.086$).

While we find that programs credentialing teachers who are more effective in math are generally also credentialing teachers who are more effective in reading, there is clear evidence throughout that training program indicators are far more predictive of teacher effectiveness in reading than math. This may suggest that more of teacher training tends to focus on the skills that teachers need to teach students reading than math.

There is no doubt that evaluating teacher training programs based on the value-added estimates of the teachers they credential is controversial. It is true that the value-added program estimates do not provide any direct guidance on how to improve teacher preparation programs. However, it is conceivable that it is not possible to move policy toward explaining *why* we see these program estimates until those estimates are first quantified. Moreover, it is certainly the case that some of the policy questions that merit investigation–-e.g., Do we see program change with new initiatives?; Do we see differences between programs within endorsement area?; How much of the difference between programs is driven by selection versus training?—require additional data. Some of these questions could be addressed with larger samples—in this case it is merely a matter of time—but other questions require additional information about individual programs, teacher candidate selection processes, and so forth. The collection of this kind of data, along with systematic analysis, would provide a path towards evidence-based reform of the pre-service portion of the teacher pipeline.

References

Aaronson, D., Barrow, L. & Sanders, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. Journal of Labor Economics, 25(1), 95-135.

Bechtold, S. & DeWitt, S. (1988). Optimal Work-Rest Scheduling with Exponential Work-Rate Decay. Management Science. 4:547-552.

Borjas, George J., & Glenn Sueyoshi, T. (1994). A two-stage estimator for probit models with structural group effects. Journal of Econometrics 64:165-182.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher Preparation and Student Achievement. Educational Evaluation and Policy Analysis, 31(4): 416-440.

Branch, G., Hanushek, E., & Rivkin, S. (2012). Estimating the Effect of Leaders on Public Sector Productivity: the Case of School Principals. CALDER Working Paper 66, January 2012.

Clotfelter, C.T., Glennie, E., Ladd, H.F., & Vigdor, J.L. (2006). Teacher Bonuses and Teacher Retention in Low Performing Schools: Evidence From the North Carolina $1,800 Teacher Bonus Program. Public Finance Review, *36* (1): 63–87.

Clotfelter, C., Ladd, H., & Vigdor, J. (2006) Teacher-Student Matching and the Assessment of Teacher Effectiveness. The Journal of Human Resources, vol.41, no. 4 (Fall 2006), p. 778-820.

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education. Washington, D.C.: The American Educational Research Association.

Crowe, E. (2010). Measuring what matters: A stronger accountability model for teacher education. Washington, DC: Center for American Progress.

Darling-Hammond, L., Wise, A., & Klein. S. (1999). A License to Teach: Raising Standards for Teaching. San Francisco, CA: Jossey-Bass.

Duncan, A. (2010). Teacher preparation: Reforming the uncertain profession Education Digest 75 (5): 13-22.

Glazerman, S., Mayer, D., & Decker, P. (2006) Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes. Journal of Policy Analysis and Management, Vol. 25, No. 1, 75–96 (2006).

Goldhaber, D. (2007). Everybody's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness. Journal of Human Resources, 42(4): 765-794.

Goldhaber, D., & Brewer, D. (2000) Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. Educational Evaluation and Policy Analysis. Vol. 22, No. 2 (Summer, 2000), pp. 129-145.

Goldhaber, D., & Hansen, M. (forthcoming). Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. Economica.

Goldhaber, D., & Liddle, S. (2011). The Gateway to the Profession. Assessing Teacher Preparation Programs Based on Student Achievement. CEDR Working Paper 2011-2. University of Washington, Seattle, WA.

Hanushek, E.A., Rivkin, S.G. & Taylor, L.L. (1996). Aggregation and the estimated effects of school resources. National Bureau of Economic Research.

Harris, D. & Sass, T. (2007). Teacher training, teacher quality and student achievement. CALDER Working Paper No. 3. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research.

Henry, G. T., Thompson, C. L., Bastian, K. C., Kershaw, D. C., Purtell, K. M., & Zulli, R. A. (2011) Does teacher preparation affect student achievement? Chapel Hill, NC: Carolina Institute for Public Policy Working Paper, Version dated February 7, 2011.

Jackson, C. K., & Bruegmann E. (2009). Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. American Economic Journal: Applied Economics 1:85-108.

Jackson, C. K. (2010). Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence From Teachers. NBER Working Paper No. 15990

Karro, J., Peifer, M., Hardison, R., Kollmann, M., & Grunberg, H. (2008). Exponential Decay of GC Content Detected by Strand-Symmetric Substitution Rates Influences the Evolution of Isochore Structure. Molecular Biology and Evolution. 25(2):362-374.

Koedel, C., & Betts. J.R. (2007). Re-examining the role of teacher quality in the educational production function. San Diego, CA: University of Missouri.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012). Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs? University of Missouri Department of Economics Working Paper Series. http://economics.missouri.edu/working-papers/2012/WP1204_koedel_et_al.pdf

Levine, A. (2006). Educating School Teachers. Washington, D.C.: The Education Schools Project.

McCaffrey, D., Sass, T., Lockwood, J.R., & Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates, Education Finance and Policy, 4 572-606.

Mihaly, K., McCaffrey, D., Sass, T., & Lockwood, J.R. (2011). Where You Come From or Where You Go? Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. RAND Working Paper, March 18, 2011

National Council for Accreditation of Teacher Education (2010). Transforming teacher education through clinical practice: A national strategy to prepare effective teachers. Washington, DC.

National Council on Teacher Quality. (2007). State Teacher Policy Yearbook: Washington State Summary. Washington, DC. Accessed 3/27/09 at http://www.nctq.org/stpy/reports/stpy_washington.pdf.

National Research Council. (2010). Better Data on Teacher Preparation Could Aid Efforts to Improve Education. The National Academies Office of News and Public Information. April 29. Accessed from: http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=12882

Nichols, A. (2008). fese: Stata module calculating standard errors for fixed effects. http://ideas.repec.org/c/boc/bocode/s456914.html

Noell, G. H., Porter, B. A., Patt, R. M. & Dahir, A. (2008). Value-added assessment of teacher preparation in Louisiana: 2004-2005 to 2006-2007. Retrieved from: http://www.laregentsarchive.com/Academic/TE/2008/Final%20Value-Added%20Report%20(12.02.08).pdf

Obizhaeva, A., & Wang, J. 2005, Optimal trading strategy and supply/demand dynamics, forthcoming Journal of Financial Markets.

Padmanabhan, G. & Vrat, P. (1990). An EOQ model for items with stock dependent consumption rate and exponential decay. Engineering Costs and Production Economics. 18(3):241.

Perlmutter, D. (2005). We Want Change; No We Don't. The Chronicle of Higher Education,

October 25, 2005.
Rivkin, S., Hanushek, E., & Kane, T. J. (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. American Economic Review, 94(2), 247-252.
Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. Quarterly Journal of Economics, 25:1.
Rubin, D., Stuart, E. and Zanutto, E. (2004). A Potential Outcomes View of Value-Added Assessment in Education. Journal of Educational and Behavioral Statistics, March 20, 2004, 29: 67-101.
Sawchuk, S. (2012). Deadlocked Negotiators Fail to Reach Consensus on Teacher-Prep Rules. Education Week Teacher Beat, April 12, 2012.
Summers, L.H. (2012) What You (Really) Need to Know. The New York Times, Jan 20, 2012.
Todd, P.E., & K.I. Wolpin. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. Economic Journal 113, F3-F33.
The Teaching Commission. (2006). Teaching at Risk: Progress & Potholes. New York, NY.
U.S. Department of Education, Office of Postsecondary Education, Preparing and Credentialing the Nation's Teachers: The Secretary's Eighth Report on Teacher Quality; Based on Data Provided for 2008, 2009 and 2010, Washington, D.C., 2011
Wennmalm, S. and Sanford, S. (2007). Studying Individual Events in Biology. Annual Review of Biochemistry. 76:419-446.

VI. Tables and Figures.

Table 1. Means of Selected Student Characteristics in 2008-09 for Unrestricted and Restricted Samples

	Unrestricted	Restricted	Difference
Math WASL Standardized Scale Score	-0.02	0.02	0.04***
Reading WASL Standardized Scale Score	-0.02	0.02	0.04***
Female Students (%)	48.2	49.1	-0.9**
American Indian Students (%)	2.4	2.3	0.1
Asian or Pacific-Islander Students (%)	9.8	8.6	1.2***
Black Students (%)	6.6	6.1	0.5**
Hispanic Students (%)	16.5	15.1	1.5***
White Students (%)	60.1	64.3	-4.2***
Multi-Racial Students (%)	4.1	3.3	0.8***
Migrant Students (%)	1.7	1.2	-0.5***
Homeless Students (%)	1.8	1.4	0.4***
English-speaking Students (%)	81.3	84.0	-2.7***
Students Eligible for Free or Reduced-Price Lunch (%)	47.4	44.7	2.7***
Students with Learning Disabilities (%)	5.8	6.1	-0.4*
Gifted Students (%)	4.9	3.6	1.3***
LEP Students (%)	9.9	6.8	3.1***
Special Education Students (%)	14.0	11.8	2.2***
Total Number of Students	53,177	58,865	-

Students in the unrestricted sample could not be matched to unique proctors. Statistical significance is denoted with: * if $p<0.05$, ** if $p<0.01$ **, and *** if $p<0.001$.

Table 2: Coefficients from 1st-Stage Value-Added Regression		
	MATH	READING
Math Pre-Score	0.581***	0.2741***
	-(0.001)	-(0.002)
Reading Pre-Score	0.177***	0.4232***
	-(0.001)	-(0.002)
Female	-0.046***	0.086***
	-(0.002)	-(0.002)
American Indian/Alaskan Native	-0.066***	-0.058***
	-(0.006)	-(0.007)
Asian/Pacific Islander	0.075***	0.021***
	-(0.003)	-(0.004)
Black	-0.105***	-0.019***
	-(0.004)	-(0.005)
Hispanic	-0.036***	-0.022***
	-(0.003)	-(0.004)
Learning Disability	0.008	-0.072***
	-(0.005)	-(0.006)
Gifted	0.263***	0.163***
	-(0.006)	-0.006
Limited English Proficiency	-0.056***	-0.169***
	-0.005	-(0.005)
Special Education	-0.207***	-0.244***
	-0.005	-0.005
Free/Reduced Price Lunch	-0.076***	-0.083***
	-0.002	-0.002
Number of Observations	397,268	397,268
R-squared	0.70	0.61

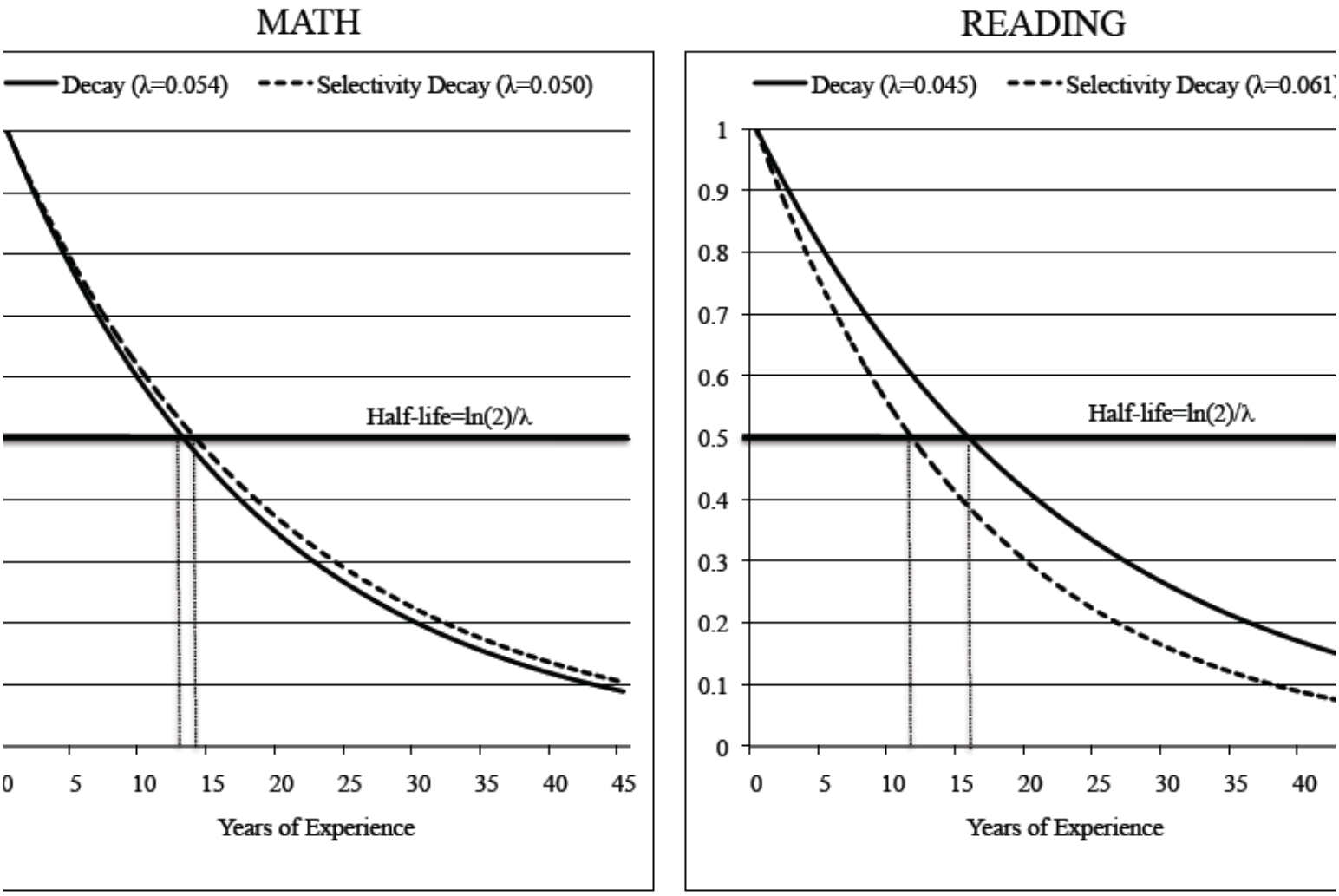
All models also include grade and year dummies. The student scores are standardized by grade and year, so all coefficients are in standard deviations of student performance. Statistical significance is denoted with: * if p<0.05, ** if p<0.01, *** if p<0.001.

Table 3. Program Estimates and Standard Errors for Various Model Specifications

	MATH			READING		
	(1)	(2)	(3)	(4)	(5)	(6)
Model Specification	No Decay	Decay	Selectivity Decay	No Decay	Decay	Selectivity Decay
Includes pre-service selectivity measures	No	No	Yes	No	No	Yes
Lambda (λ)	NA	0.054	0.050†	NA	0.045†	0.061**
	NA	(0.04)	(0.03)	NA	(0.02)	(0.02)
Antioch	-4.43*	-5.05†	-5.22†	-0.05	-0.19	-0.35
	(2.26)	(3.02)	(2.96)	(1.92)	(2.40)	(2.57)
Central Washington	-1.01	-0.52	-3.23	0.37	0.4	-1.03
	(0.78)	(1.29)	(2.04)	(0.72)	(1.11)	(1.79)
City	-1.28	-0.67	-0.97	-0.47	-0.52	-0.63
	(1.09)	(1.36)	(1.36)	(1.00)	(1.24)	(1.33)
Eastern Washington	-0.63	0.35	-2.46	-2.37**	-3.86*	-5.71**
	(0.95)	(1.59)	(2.11)	(0.85)	(1.57)	(2.23)
Gonzaga	1.49	4.57	2.35	-3.00†	-4.33†	-7.59*
	(2.11)	(3.18)	(3.40)	(1.69)	(2.50)	(3.23)
Heritage	-0.02	1.08	0.78	-0.96	-1.13	-1.19
	(1.89)	(2.55)	(2.51)	(1.73)	(2.22)	(2.39)
Northwest	-5.33	-6.84	-6.78	-1.12	-3.44	-3.99
	(3.75)	(4.82)	(4.98)	(3.82)	(5.21)	(5.87)
Pacific Lutheran	0.83	2.19	-0.29	-1.98†	-4.38	-8.42**
	(1.28)	(2.13)	(2.52)	(1.21)	(1.90)	(2.80)
St. Martin's	-2.66†	-3.29	-6.07*	-5.67***	-8.89***	-11.03***
	(1.63)	(2.56)	(3.03)	(1.64)	(2.62)	(3.24)
Seattle Pacific	0.15	1.61	-1.32	0.86	2.2	-0.66
	(1.69)	(2.75)	(3.10)	(1.43)	(2.28)	(2.91)
Seattle U	-2.38	-0.48	-2.53	-0.70	0.49	-0.81
	(1.64)	(2.51)	(2.87)	(1.47)	(2.13)	(2.73)
Evergreen	-4.70	-6.5	-9.88†	-2.07	-3.18	-4.00
	(3.45)	(5.24)	(6.13)	(3.95)	(5.50)	(6.66)
U of Puget Sound	1.03	3.87	3.15	1.74	2.78	2.08
	(1.76)	(3.46)	(3.83)	(1.52)	(2.79)	(3.71)
UW Seattle	1.39	4.7*	3.56	1.06	2.15	0.47
	(1.15)	(2.37)	(2.67)	(1.00)	(1.66)	(2.38)
UW Bothell	0.91	2.26	1.89	4.57**	5.57**	5.76**
	(2.09)	(2.55)	(2.53)	(1.75)	(2.16)	(2.29)
UW Tacoma	2.00	3.12	2.79	2.92	3.23	3.14
	(2.10)	(2.73)	(2.69)	(1.89)	(2.38)	(2.54)
Walla Walla	0.91	1.06	0.63	7.11*	10.32†	11.00†
	(6.38)	(11.34)	(11.04)	(3.49)	(5.87)	(6.50)
Washington State	0.08	0.64	-1.81	0.72	0.81	-1.31
	(0.89)	(1.29)	(1.80)	(0.80)	(1.10)	(1.73)
Western Washington	-0.38	0.92	-1.12	1.14	1.86†	-1.29
	(0.82)	(1.24)	(1.95)	(0.75)	(1.14)	(1.95)
Whitworth	1.25	3.82	1.18	-1.23	-0.45	-2.86
	(1.52)	(2.41)	(2.69)	(1.34)	1.88*	(2.60)

Reported program coefficients and standard errors are multiplied by 100. All models include teacher and classroom covariates (i.e., gender, race/ethnicity, degree level, experience, class size). Selectivity Decay models also include measures of selectivity (i.e., freshman admissions rates, composite SAT scores, and percent freshman with GPA > 3.0). Statistical significance: † if $p < 0.10$, * if $p < 0.05$, ** if $p < 0.01$, and *** if $p < 0.001$.

Figure 1. Decay Curves and Half-life for Decay and Selectivity Decay Models



Y-axis = $e(-\lambda * Experience)$

Table 4. Program Estimates From Decay and Robustness Models

	MATH			READING		
	(1)	(2)	(3)	(4)	(5)	(6)
Model Specification	Decay	District FE w/ Decay	School FE w/ Decay	Decay	District FE w/ Decay	School FE w/ Decay
Lambda (λ)	0.054	-0.037*	-0.044	0.045†	0.013	-0.015
	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Antioch	-5.05†	-2.81	-2.48	-0.19	0.14	1.02
	(3.02)	(1.81)	(1.83)	(2.40)	(2.18)	(1.77)
Central Washington	-0.52	-0.44	-0.16	0.4	0.58	0.99
	(1.29)	(0.45)	(0.34)	(1.11)	(0.90)	(0.62)
City	-0.67	-0.41	-0.3	-0.52	-0.53	0.01
	(1.36)	(0.91)	(0.80)	(1.24)	(1.08)	(0.83)
Eastern Washington	0.35	-1.77*	-1.36†	-3.86*	-3.28*	-2.28*
	(1.59)	(0.89)	(0.71)	(1.57)	(1.33)	(0.94)
Gonzaga	4.57	-2.01	-1.68	-4.33†	-4.52*	-3.69*
	(3.18)	(1.40)	(1.22)	(2.50)	(2.02)	(1.65)
Heritage	1.08	-0.35	-0.09	-1.13	-2.06	-0.5
	(2.55)	(1.39)	(1.19)	(2.22)	(1.92)	(1.54)
Northwest	-6.84	-2.65	1.96	-3.44	-1.02	2.84
	(4.82)	(2.61)	(2.28)	(5.21)	(4.73)	(2.96)
Pacific Lutheran	2.19	0.29	0.61	-4.38	-2.8	-0.82
	(2.13)	(0.66)	(0.59)	(1.90)	(1.52)	(0.97)
St. Martin's	-3.29	-0.41	-0.6	-8.89***	-3.39†	-2.24
	(2.56)	(1.12)	(0.99)	(2.62)	(1.97)	(1.57)
Seattle Pacific	1.61	-0.21	-0.38	2.2	1.23†	1.08
	(2.75)	(0.84)	(0.73)	(2.28)	(1.62)	(1.13)
Seattle U	-0.48	-2.28†	-1.99†	0.49	-2.08	-0.77
	(2.51)	(1.26)	(1.12)	(2.13)	(1.75)	(1.21)
Evergreen	-6.5	-2.84	-2.06	-3.18	-2.85	-2.35
	(5.24)	(2.41)	(2.11)	(5.50)	(3.92)	(2.98)
U of Puget Sound	3.87	-0.05	0.29	2.78	1.33	1.57
	(3.46)	(0.82)	(0.62)	(2.79)	(1.98)	(1.22)
UW Seattle	4.7*	-0.48	-0.58	2.15	-0.54	-0.25
	(2.37)	(0.58)	(0.51)	(1.66)	(1.18)	(0.71)
UW Bothell	2.26	0.48	0.62	5.57**	2.01	2.33
	(2.55)	(1.84)	(1.76)	(2.16)	(1.96)	(1.69)
UW Tacoma	3.12	1.04	-0.11	3.23	2.55	0.33
	(2.73)	(1.73)	(1.64)	(2.38)	(2.08)	(1.61)
Walla Walla	1.06	0.55	-1.33	10.32†	7.66	5.55
	(11.34)	(3.78)	(3.79)	(5.87)	(4.98)	(3.89)
Washington State	0.64	-0.28	-0.29	0.81	-0.06	0.12
	(1.29)	(0.53)	(0.42)	(1.10)	(0.88)	(0.60)
Western Washington	0.92	-0.69	-0.72	1.86†	0.19	0.11
	(1.24)	(0.56)	(0.49)	(1.14)	(0.87)	(0.57)
Whitworth	3.82	-1.38	-1.64	-0.45	-1.44	-1.52
	(2.41)	(1.05)	(1.02)	1.88*	(1.62)	(1.19)

Reported program coefficients and standard errors are multiplied by 100. All models include teacher and classroom covariates (i.e., gender, race/ethnicity, degree level, experience, class size). District FE Decay models include district fixed effects. School FE Decay models include school fixed effects. Statistical significance: † if p <0.10, * if p<0.05, ** if p<0.01, and *** if p<0.001.

Table 5. Pearson | Spearman Correlations of Program Estimates from Various Model Specifications

		MATH					READING				
	Model Specification	No Decay	Decay	Selectivity Decay	District FE Decay	School FE Decay	No Decay	Decay	Selectivity Decay	District FE Decay	School FE Decay
MATH	No Decay	1|1									
	Decay	0.97|0.96	1|1								
	Selectivity Decay	0.92|0.92	0.95|0.94	1|1							
	District FE Decay	0.73|0.58	0.60|0.48	0.63|0.56	1|1						
	School FE Decay	0.11|0.17	0.05|0.19	0.16|0.28	0.45|0.61	1|1					
READING	No Decay	0.41|0.40	0.32|0.33	0.47|0.45	0.53|0.60	0.15|0.32	1|1				
	Decay	0.43|0.45	0.38|0.38	0.50|0.46	0.48|0.53	0.02|0.16	0.98|0.96	1|1			
	Selectivity Decay	0.31|0.34	0.24|0.28	0.41|0.46	0.42|0.47	0.03|0.17	0.96|0.90	0.97|0.90	1|1		
	District FE Decay	0.29|0.31	0.17|0.20	0.32|0.32	0.55|0.60	0.14|0.36	0.93|0.95	0.91|0.90	0.91|0.86	1|1	
	School FE Decay	0.02|0.04	-0.07|-0.03	0.13|0.09	0.37|0.44	0.39|0.49	0.82|0.79	0.76|0.70	0.79|0.72	0.89|0.87	1|1

Table 6. Main and Interaction Effects Between Program Indicators and Graduation Date

	MATH			READING		
	(1)	(2)	(3)	(4)	(5)	(6)
	Main Effects	2000-2004	2005-2009	Main Effects	2000-2004	2005-2009
Antioch	0.031	-0.088†	-0.08	0.043	-0.037	-0.048
	(0.039)	(0.051)	(0.057)	(0.034)	(0.046)	(0.047)
Central	-0.012	-0.011	0.044†	-0.006	0.022	0.049**
	(0.010)	(0.020)	(0.023)	(0.009)	(0.019)	(0.019)
City	-0.017	0.008	0.022	-0.005	-0.011	0.036
	(0.021)	(0.028)	(0.027)	(0.019)	(0.026)	(0.024)
Eastern	-0.009	0.004	0.032	-0.029***	0.008	0.054†
	(0.011)	(0.024)	(0.031)	(0.010)	(0.022)	(0.030)
Heritage	-0.021	0.031	0.055	-0.013	-0.028	0.072†
	(0.030)	(0.043)	(0.045)	(0.028)	(0.039)	(0.040)
Pacific Lutheran	0.004	0.027	0.003	-0.017	-0.01	-0.006
	(0.015)	(0.035)	(0.033)	(0.015)	(0.030)	(0.032)
Seattle U	-0.055	0.078*	0.098**	-0.033†	0.071*	0.069†
	(0.020)	(0.039)	(0.044)	(0.020)	(0.033)	(0.038)
UW Seattle	0.007	0.048	0.011	0.009	-0.025	0.028
	(0.014)	(0.031)	(0.035)	(0.012)	(0.029)	(0.027)
UW Bothell	-0.031	0.045	0.067	-0.056	0.118**	0.137**
	(0.060)	(0.067)	(0.070)	(0.040)	(0.046)	(0.050)
Washington State	0.001	0.018	-0.01	0.003	0.009	0.031
	(0.012)	(0.021)	(0.023)	(0.011)	(0.019)	(0.020)
Western	-0.011	0.021	0.029	0.006	0.012	0.037†
	(0.010)	(0.020)	(0.024)	(0.009)	(0.018)	(0.021)
Whitworth	0.004	0.031	0.024	-0.036*	0.064*	0.084**
	(0.020)	(0.035)	(0.042)	(0.018)	(0.030)	(0.034)
Other	-0.006	0.006	0.036	-0.01	0.009	0.025
	(0.011)	(0.022)	(0.024)	(0.010)	(0.020)	(0.023)

Interaction effects are only reported for the twelve institutions with at least 30 graduates from both time periods. All other programs are collapsed and their combined effects are reported together in the "other" category. Coefficients for the 2000-2004 and 2005-2009 interactions should be interpreted relative to the main effect (representing graduates before 2000).

Both math and reading models also include teacher covariates (gender, race/ethnicity, degree level, and experience) and classroom covariates (aggregated student demographic characteristics and class size).

Statistical significance is denoted with: † if $p < 0.10$, * if $p < 0.05$, ** if $p < 0.01$, and *** if $p < 0.001$.

Appendix

Table A1. Mean Teacher Characteristics of Teachers from Different Training Programs							
	Out of State	Antioch University	Central Washington University	City University	Eastern Washington University	Evergreen State College	Gonzaga University
Number of Teachers	1891	82	1181	534	753	49	119
Female (%)	81.3	72.0	78.4	69.0	73.9	71.4	84.9
American Indian (%)	0.5	0.0	0.4	0.9	0.3	2.0	0.8
Asian Pacific Islander (%)	2.9	11.0	1.8	3.4	1.5	6.1	3.1
Black (%)	1.4	8.5	0.8	3.9	0.1	0.0	0.0
Hispanic (%)	1.7	1.2	2.5	2.3	2.8	0.0	0.8
White (%)	93.5	79.3	94.4	89.5	95.3	89.8	95.2
Masters Degree or higher (%)	65.5	84.1	53.9	88.2	68.7	100.0	74.6
Average Teaching Experience	14.2	4.6	13.2	5.1	14.2	7.4	12.3
< 1 Yr Teaching Experience (%)	2.8	10.1	4.7	10.7	2.1	13.3	4.6
1-2 Yrs Teaching Experience (%)	3.4	13.3	5.2	12.8	4.3	6.1	3.8
2-3 Yrs Teaching Experience (%)	3.4	13.9	5.7	14.4	4.5	6.8	5.3
3-4 Yrs Teaching Experience (%)	4.8	9.8	4.5	11.3	4.5	3.1	5.3
4-5 Yrs Teaching Experience (%)	4.9	11.0	3.6	9.4	3.1	8.8	5.9
5+ Yrs Teaching Experience (%)	80.8	42.0	76.3	41.4	81.5	61.9	75.1
	Heritage University	Northwest University	Pacific Lutheran University	Saint Martin's University	Seattle Pacific University	Seattle University	University of Puget Sound
Number of Teachers	179	26	345	147	199	164	165
Female (%)	77.7	72.0	81.4	78.9	79.4	75.5	75.0
American Indian (%)	1.1	0.0	0.9	1.4	0.0	1.2	1.2
Asian Pacific Islander (%)	2.2	4.0	2.9	1.4	1.5	6.1	1.8
Black (%)	2.8	0.0	0.3	2.7	1.5	2.5	0.6
Hispanic (%)	23.5	0.0	1.4	1.4	2.0	1.2	3.0
White (%)	70.4	96.0	94.5	93.2	95.0	89.0	93.3
Masters Degree or higher (%)	67.6	50.0	66.4	54.7	54.8	86.8	75.0
Average Teaching Experience	6.3	7.8	12.3	9.8	13.3	11.7	16.5
< 1 Yr Teaching Experience (%)	8.3	13.5	5.0	8.0	6.6	6.1	4.3
1-2 Yrs Teaching Experience (%)	11.8	15.4	6.4	7.7	4.8	6.7	3.5
2-3 Yrs Teaching Experience (%)	9.1	3.2	4.6	3.3	2.8	7.8	3.1
3-4 Yrs Teaching Experience (%)	11.6	8.3	5.0	3.5	1.7	5.4	1.4
4-5 Yrs Teaching Experience (%)	6.9	1.3	5.2	2.9	1.5	4.7	1.6
5+ Yrs Teaching Experience (%)	52.3	58.3	73.8	74.5	82.6	69.3	86.2

Table A1. (continued)

	University of Washington Seattle	University of Washington Bothell	University of Washington Tacoma	Walla Walla University	Washington State University	Western Washington University	Whitworth University
Number of Teachers	474	105	89	14	915	1050	237
Female (%)	76.8	81.0	79.5	71.4	81.8	79.6	74.9
American Indian (%)	1.1	1.0	0.0	0.0	0.6	0.9	0.6
Asian Pacific Islander (%)	7.6	2.9	5.7	0.0	1.2	2.0	3.4
Black (%)	1.1	1.0	3.4	0.0	0.9	0.8	0.9
Hispanic (%)	2.7	5.7	1.1	0.0	3.0	0.9	1.3
White (%)	87.6	89.5	89.8	100.0	94.3	95.5	93.8
Masters Degree or higher (%)	64.0	31.5	26.1	82.1	61.9	53.2	71.9
Average Teaching Experience	16.9	4.1	5.0	11.5	11.7	12.5	11.5
< 1 Yr Teaching Experience (%)	3.2	12.0	11.3	0.0	6.5	4.1	5.0
1-2 Yrs Teaching Experience (%)	3.3	14.1	11.4	0.0	7.8	5.2	7.5
2-3 Yrs Teaching Experience (%)	3.0	15.8	11.9	7.1	6.9	4.7	6.4
3-4 Yrs Teaching Experience (%)	2.9	10.2	7.0	0.0	6.9	5.1	6.5
4-5 Yrs Teaching Experience (%)	4.0	4.8	5.2	0.0	4.6	4.2	5.0
5+ Yrs Teaching Experience (%)	83.5	43.0	53.1	92.9	67.2	76.7	69.5

Table A2. Mean Student Characteristics of Teachers from Different Training Programs							
	Out of State	Antioch University	Central Washington University	City University	Eastern Washington University	Evergreen State College	Gonzaga University
Number of Students	81595	3457	51816	22723	34461	1943	5909
Math score (mean)	406.3	401.1	400.8	405.1	404.5	401.7	407.7
Math score (std dev)	41.4	40.3	40.6	41.0	40.8	42.2	41.7
Reading score (mean)	411.2	408.9	409.0	410.5	409.8	409.4	410.5
Reading score (std dev)	23.65	23.47	23.38	23.78	24.66	25.00	24.66
Female (%)	49.2	48.9	49.0	48.9	49.0	49.2	48.0
American Indian (%)	2.3	1.8	2.4	2.2	3.0	3.7	3.1
Asian Pacific Islander (%)	8.8	18.0	7.0	10.5	4.6	11.8	5.3
Black (%)	5.3	12.6	5.2	7.1	3.8	9.9	3.4
Hispanic (%)	13.2	13.8	23.5	11.5	13.8	11.0	9.4
White (%)	67.6	51.2	59.1	65.3	71.9	61.8	75.0
Multiracial (%)	2.5	2.3	2.2	3.1	2.7	1.3	3.6
Free Lunch (%)	39.8	44.9	46.9	37.9	47.7	42.4	43.4
Learning Disabled (%)	7.9	8.6	8.0	8.3	7.8	9.1	8.2
Gifted (%)	4.8	2.1	4.0	3.9	4.4	4.9	4.4
LEP (%)	5.2	6.9	7.6	4.5	4.9	4.4	3.5
Special Education (%)	11.7	12.7	11.6	12.3	11.7	13.1	12.8
	Heritage University	Northwest University	Pacific Lutheran University	Saint Martin's University	Seattle Pacific University	Seattle University	University of Puget Sound
Number of Students	6901	950	15861	6525	8891	6886	8170
Math score (mean)	389.9	404.4	404.9	399.4	410.3	410.9	406.4
Math score (std dev)	40.4	42.4	40.1	39.4	42.8	42.4	40.2
Reading score (mean)	403.2	410.6	410.6	407.6	413.0	413.4	411.5
Reading score (std dev)	23.2	23.6	23.3	23.6	23.4	23.1	23.4
Female (%)	49.3	49.6	49.7	49.8	48.9	49.1	49.0
American Indian (%)	4.6	1.6	2.3	3.0	1.6	1.1	2.6
Asian Pacific Islander (%)	2.7	14.8	9.5	6.5	11.6	17.0	10.3
Black (%)	3.0	6.2	8.4	7.2	5.9	7.5	9.6
Hispanic (%)	56.2	13.2	10.8	12.0	13.2	12.5	10.9
White (%)	32.7	59.9	66.5	68.4	64.7	58.8	63.7
Multiracial (%)	0.8	4.2	2.2	2.2	2.6	2.9	2.4
Free Lunch (%)	70.5	35.2	38.0	49.3	34.8	34.1	39.5
Learning Disabled (%)	7.6	8.2	7.0	7.8	8.0	8.9	7.7
Gifted (%)	2.4	5.9	3.8	3.7	6.8	5.6	4.3
LEP (%)	17.2	5.4	3.2	3.2	4.7	5.8	3.5
Special Education (%)	9.7	10.9	11.2	12.1	12.0	12.3	11.5

Table A2. (continued)

	University of Washington Seattle	University of Washington Bothell	University of Washington Tacoma	Walla Walla University	Washington State University	Western Washington University	Whitworth University
Number of Students	22342	5002	3877	679	40871	49230	10581
Math score (mean)	412.7	410.0	397.6	394.2	404.9	405.7	405.7
Math score (std dev)	43.0	43.4	39.2	40.8	41.1	40.4	40.8
Reading score (mean)	414.1	413.0	407.7	406.5	410.7	411.3	409.8
Reading score (std dev)	23.6	24.4	22.4	23.8	23.8	23.6	25.3
Female (%)	49.4	49.6	47.7	51.3	49.3	49.1	49.4
American Indian (%)	1.7	1.8	1.6	1.9	1.9	3.0	2.9
Asian Pacific Islander (%)	13.9	15.8	12.0	4.3	6.8	7.9	4.3
Black (%)	6.7	5.2	19.3	2.5	3.9	4.1	4.1
Hispanic (%)	12.3	13.4	12.7	25.7	17.6	12.8	9.8
White (%)	62.3	59.6	52.4	65.5	67.0	69.6	75.6
Multiracial (%)	2.8	4.1	1.7	0.1	2.5	2.2	3.1
Free Lunch (%)	32.1	34.9	52.3	54.8	42.8	39.5	46.4
Learning Disabled (%)	8.3	8.8	7.6	9.6	7.6	8.3	8.2
Gifted (%)	5.4	4.7	2.7	3.1	3.8	3.5	4.6
LEP (%)	4.7	6.4	4.5	12.7	6.5	4.8	3.8
Special Education (%)	12.4	12.4	10.8	13.6	11.4	12.5	12.4

Table A3. Student-to-Teacher Matching by Subject, Grade, and Year

		MATH					READING				
		Grade 3	Grade 4	Grade 5	Grade 6	All Grades	Grade 3	Grade 4	Grade 5	Grade 6	All Grades
2005-06	Useable matches	65,595	64,275	64,849	48,030	242,749	64,879	64,047	65,379	47,741	242,046
		87.1%	86.2%	84.8%	62.5%	80.1%	86.2%	85.9%	85.5%	62.1%	79.8%
	Unuseable matches	2,174	2,209	1,802	6,363	12,548	2,705	2,270	2,096	6,192	13,263
		2.9%	3.0%	2.4%	8.3%	4.1%	3.6%	3.0%	2.7%	8.1%	4.4%
	Unmatched	7,528	8,104	9,782	22,434	47,848	7,713	8,271	8,958	22,894	47,836
		10.0%	10.9%	12.8%	29.2%	15.8%	10.2%	11.1%	11.7%	29.8%	15.8%
	No. of Students	75,297	74,588	76,433	76,827	303,145	75,297	74,588	76,433	76,827	303,145
2006-07	Useable matches	66,244	65,465	63,881	47,918	243,508	65,764	65,365	64,159	47,750	243,038
		87.1%	86.3%	84.8%	62.5%	80.1%	86.5%	86.2%	85.2%	62.2%	80.0%
	Unuseable matches	2,399	2,182	2,190	5,235	12,006	2,758	2,407	2,474	5,486	13,125
		3.2%	2.9%	2.9%	6.8%	4.0%	3.6%	3.2%	3.3%	7.2%	4.3%
	Unmatched	7,415	8,169	9,224	23,555	48,363	7,536	8,044	8,662	23,472	47,714
		9.7%	10.8%	12.3%	30.7%	15.9%	9.9%	10.6%	11.5%	30.6%	15.7%
	No. of Students	76,058	75,816	75,295	76,708	303,877	76,058	75,816	75,295	76,708	303,877
2007-08	Useable matches	65,226	64,648	63,829	44,156	237,859	65,152	64,940	63,941	43,878	237,911
		84.1%	83.5%	82.7%	57.8%	77.1%	84.0%	83.9%	82.9%	57.4%	77.1%
	Unuseable matches	2,374	2,536	2,502	5,002	12,414	2,460	2,499	2,429	5,197	12,585
		3.1%	3.3%	3.2%	6.5%	4.0%	3.2%	3.2%	3.1%	6.8%	4.1%
	Unmatched	9,990	10,221	10,806	27,275	58,292	9,978	9,966	10,767	27,358	58,069
		12.9%	13.2%	14.0%	35.7%	18.9%	12.9%	12.9%	14.0%	35.8%	18.8%
	No. of Students	77,590	77,405	77,137	76,433	308,565	77,590	77,405	77,137	76,433	308,565
2008-09	Useable matches	33,958	33,943	33,607	22,457	123,965	33,937	33,978	33,584	22,071	123,570
		43.3%	43.2%	42.9%	28.9%	39.6%	43.2%	43.2%	42.9%	28.4%	39.5%
	Unuseable matches	1,752	1,589	1,618	1,212	6,171	1,723	1,601	1,592	1,433	6,349
		2.2%	2.0%	2.1%	1.6%	2.0%	2.2%	2.0%	2.0%	1.8%	2.0%
	Unmatched	42,780	43,054	43,043	54,011	182,888	42,830	43,007	43,092	54,176	183,105
		54.5%	54.8%	55.0%	69.5%	58.4%	54.6%	54.7%	55.1%	69.7%	58.5%
	No. of Students	78,490	78,586	78,268	77,680	313,024	78,490	78,586	78,268	77,680	313,024

Note: Unuseable matches are for proctor names that matched more than one teacher name in a given school.

Table A4. Number of Teachers with WEST-B Scores by Program		
Program Name	Count	Percent
Out of State	235	16.0
Antioch University	30	2.0
Central Washington University	156	10.6
City University	238	16.2
Eastern Washington University	62	4.2
Evergreen State College	10	0.7
Gonzaga University	12	0.8
Heritage University	59	4.0
Northwest University	9	0.6
Pacific Lutheran University	55	3.7
Seattle Pacific University	24	1.6
Seattle University	35	2.4
St. Martin's University	25	1.7
University of Puget Sound	22	1.5
University of Washington Bothell	50	3.4
University of Washington Seattle	47	3.2
University of Washington Tacoma	36	2.5
Walla Walla University	1	0.1
Washington State University	192	13.1
Western Washington University	128	8.7
Whitworth University	43	2.9
Total	1,469	100.0