Is a Good Teacher a Good Teacher for All? Comparing Value-Added of Teachers with Their English Learners and Non-English Learners

Susanna Loeb, James Soland, Lindsay Fox

Abstract

Districts, states and researchers are using value-added models with increasing frequency to evaluate educational policies and programs, as well as teachers and other educators individually. Despite their prevalence, little research assesses whether value-added measures (VAM) are consistent across student subgroups. Are teachers who are effective with one group of students also effective with others? If they are not, then it may be worthwhile to develop separate measures of teacher effectiveness for different student groups; if they are, a single average measure will likely suffice. Our paper uses data from a large urban district with a considerable English learner (EL) population to compare teachers' VAM with ELs to the same teachers' VAM with non-ELs. We find that teachers who are effective with ELs also tend to be effective with their non-ELs and vice-versa. We also, however, find evidence that some teachers are relatively more effective with ELs than with non-ELs, and that this increased efficacy is predicted by a teacher's fluency in students' home language and whether he or she possesses a bilingual teaching certification.

Keywords: value added estimation, English language learners, teacher labor markets, teacher policy

Teacher effectiveness has been the focus of much recent education reform, including the federal Race to the Top program and the Teacher Incentive Fund. Teacher quality has also been a dominant feature of much recent education research, including studies of how to measure teacher quality (Hanushek & Rivkin, 2010; Kane & Staiger, 2012), how to hire effective teachers (Rockoff, Jacob, Kane, & Staiger, 2011), and how to improve teacher effectiveness (Hill, 2007; Loeb, Kalogrides, & Beteille, 2012). A common assumption underlying these policies and research approaches is that a teacher who is effective for one group of students is also effective for other groups of students. With some exceptions, few studies have assessed the relative effectiveness of teachers with different types of students (Aaronson, Barrow, & Sander, 2007; Dee, 2005, 2007; Lockwood & McCaffrey, 2009; Loeb & Candelaria, 2012). This gap in the research occurs despite studies showing some student subgroups may benefit from specialized instructional approaches. States, districts and schools are expending substantial effort in teacher professional development for teaching English learners (ELs). These students may benefit from having teachers with EL-specific training and fluency in the student's native language (Master, Loeb, Whitney, & Wyckoff, 2012).

In this paper, we assess the extent to which a teacher's effectiveness at improving student performance in math and reading is similar for ELs and their non-EL counterparts. In particular, we ask three research questions. (1) How much does teacher effectiveness vary across classrooms for EL and non-EL students? (2) Are teachers equally effective with ELs and non-ELs? Finally, (3) can measured teacher characteristics help explain differences in teacher effectiveness?

The paper proceeds as follows. First, we motivate and focus the study drawing on literature about teacher quality and effective instruction for English language learners. We then

present the data, methods, and findings. Finally, we conclude with a discussion of the results. Overall, we find that, with some exceptions, teachers explain a similar amount of learning for EL and non-EL students. We also find that, on average, teachers who are effective with non-EL students are also effective with ELs, though some teachers are differentially effective with one group or the other. While we only touch on characteristics of teachers that explain differential learning, we find that teachers who speak the native language of ELs or possess bilingual certification tend to produce relatively greater gains for ELs than for non-ELs.

Background

Value-added estimates—the amount teachers increase the achievement test scores of their students over the course of the year—have become a popular measure of teacher effectiveness for policy makers. Though no consensus exists on the most accurate gauge of a teacher's contribution to student outcomes, value-added measures have the benefits of measuring student learning directly, being relatively low-cost to calculate for some teachers given the testing regimes already in place, and reducing many forms of bias (Rubin, Stuart, & Zanutto, 2004). This last facet of value-added is especially important given teachers are not randomly assigned to students or schools, which can conflate the influence of student, school, and teacher variables on achievement (Clotfelter, Ladd, & Vigdor, 2007; Feng, 2010). In fact, extant research provides evidence that teachers are often assigned to particular schools and classrooms based on specific characteristics, such as their experience and teaching ability (Kalogrides, Loeb, & Beteille, forthcoming). While value-added measures may not account completely for this sorting, they address the sorting more directly than do most other measures of teacher effectiveness that are collected on a large scale, such as observational measures (McCaffrey, 2012; Rothstein, 2009).

Despite the prevalence of value-added measures, value-added research often relies on a fundamental yet untested assumption: that a teacher who is effective for one student is effective for other students with different needs (Reardon & Raudenbush, 2009). To date, little research considers whether value-added is consistent across different student subgroups, such as ethnic and language minority students. This omission occurs even though studies provide evidence that teachers can have differential effects for various student subgroups, including ELs (Dee, 2005, 2007; Master, Loeb, Whitney, & Wyckoff, 2012). Exceptions to this gap in the value-added literature include studies by Aaronson, Barrow, and Sander (2007) and Lockwood and McCaffrey (2009). Both papers produce estimates for teachers serving high- and low-performing students, showing that teachers can have differential effects on the achievement of these two groups, though the differences tend to be small. Otherwise, no research (of which we are aware) produces distinct value-added estimates by subgroup. As a result, current value-added studies can help educators determine which teachers are effective, on average, for the students they serve, but may not provide useful information on which teachers are best equipped to serve specific groups of students such as low-income or other at-risk student populations most in need of effective teaching. While there may not be compelling reasons why some groups of students would be differentially served by teachers, there are compelling reasons to believe that certain populations of students may benefit from different instructional approaches. English learners and special education students are two such examples.

Our study starts to close this gap in the value-added literature by generating separate value-added estimates for EL and non-EL students. We choose ELs because they are a rapidly growing subgroup with unique educational challenges and, therefore, may benefit from EL-specific instructional strategies (Abedi, Hofstetter, & Lord, 2004; August & Pease-Alvarez,

1996; Master, Loeb, Whitney, & Wyckoff, 2012; Solomon, Lalas, & Franklin, 2006). The research documenting these challenges is abundant. English learners enter school with lower rates of math and English proficiency, and these gaps persist well into their schooling (Parrish et al., 2006; Reardon & Galindo, 2009; Rumberger & Gandara, 2004). Based on test scores from the National Assessment of Education Progress (NAEP), 71 percent of ELs remain below basic in math and Language Arts in eighth grade compared to roughly 20 percent for non-EL students (Fry, 2007). ELs also prove less likely to progress through school than any other student subgroup (Kao & Thompson, 2003). While these statistics are complicated by several factors, including requirements in some states that EL students be proficient in both basic English and Language Arts to be reclassified as fully English proficient, the academic challenges faced by ELs are no less real.

Given the educational challenges confronted by ELs, researchers have begun to consider differential teacher effectiveness with ELs. Though most research on effective educational practices for ELs has focused on programmatic aspects of instruction (August & Shanahan, 2007; Slavin & Cheung, 2005; Tellez & Waxman, 2006), some research has addressed teaching practices for teachers of English learners (Abedi, Hofstetter, & Lord, 2004; Solomon, Lalas, & Franklin, 2006). For example, Master, Loeb, Whitney, & Wyckoff (2012) explored whether ELs benefit differentially in terms of math learning from having teachers with particular characteristics such as prior experience teaching English learners. This research builds on prior studies showing that, in some cases (though not all), teachers with more than a year or two of experience (Clotfelter, Ladd, & Vigdor, 2007; Harris & Sass, 2011; Kane, Rockoff, & Staiger, 2008; Nye, Konstantopoulos, & Hedges, 2004; Rice, 2003; Wayne & Youngs, 2003), specific content knowledge (Hill, Rowan, & Ball, 2005; Rockoff, Jacob, Kane, & Staiger, 2011), and

particular types of preparation (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Ronfeldt, 2012) can be more effective. Finally, some research finds that ELs tend to learn more in schools with practices designed to increase the effectiveness of teachers with ELs, though these results are only suggestive of an effect (Williams et al., 2007). In total, this body of research supports the contention that particular teacher skills may influence EL achievement, and that schools can adopt practices that may help their teachers develop these skills.

In the remainder of this paper, we model value-added for teachers of ELs and non-ELs to help determine whether some teachers are differentially effective with these groups and, if so, which teacher characteristics predict differential effectiveness. Our findings, in turn, help answer our underlying research question: is an effective teacher for students fluent in English also effective for ELs?

Data

We use data from the Miami-Dade County Public Schools (M-DCPS) district from the 2004-05 through 2010-11 school years. Nationwide, M-DCPS is the fourth largest school district and has a considerable EL population. In 2010-11, there were over 347,000 students enrolled in 435 schools. Of those students, more than 225,000 were Hispanic and more than 67,000 were ELs. In addition to its size and large EL population, Miami is well suited for our study because teachers transfer out of the district at relatively low rates, which provides a stable cohort for value-added analysis. Due in part to this stability and large sample size, we are able to estimate value-added by grade level, which acknowledges the different educational needs that ELs may have at various stages of their schooling.

For all of our EL value-added estimates, we use two definitions of "English learner" to identify students for the analysis. First, we use the definition of EL in the M-DCPS

administrative dataset, which follows state and federal law, as well as local policy requirements. This first definition is, in essence, the one most educators and policymakers use when they consider a student to be an English learner.[1] One problem with this definition is that ELs recently reclassified as non-ELs may still benefit from similar instructional practices as ELs. Given many of the non-ELs in the dataset who share classes with ELs may fall into this just-reclassified group, our comparison might be weighted toward a contrast between ELs and those recently reclassified rather than students who were always non-EL. Therefore, we run the risk of failing to identify the true differences in instructional needs between ELs and fully English proficient students. To supplement the first definition of EL, we use a second approach in which we define ELs as any student who either is identified as such in the data, or who was classified as an EL within the past three years. This approach is similar to a federal policy that addresses the same issue by allowing states to count students reclassified as fully English proficient in the EL cohort for two years after exiting EL status. It reduces the problem of having our analyses based on comparisons of ELs and just reclassified non-ELs, though it will rely partially on comparisons within the non-EL group of those reclassified and never-EL students. Given the complexities of classifying students as proficient or not proficient in English, we do not privilege either definition. Rather, we use each as a robustness check for the other. Given our results are similar regardless of the strategy used, we focus primarily on a single definition, the second one, but also provide the main results for the first definition.

To construct our analytic data file, we combine several datasets. First, we obtain demographic data on students from an administrative database that includes race, gender, free or

[1] This definition matches the one used by M-DCPS and is therefore reflected in the administrative data used in our sample. In the district, an EL in grades 3-9 can be reclassified as non-EL or "Fully English Proficient" if he or she meets the following criteria (at minimum): (1) performs at grade level on the overall CELLA, (2) scores proficient or above on the CELLA listening/speaking subtests, (3) scores proficient or above on the CELLA writing subtest, and (4) earns a passing score on the FCAT reading test (3 or higher).

reduced-price lunch eligibility, special education status, and whether the students are limited English proficient. Second, we combine the demographic data with test score data in order to calculate achievement gains in math and reading for students in a given teacher's classroom. The test score data come from the Florida Comprehensive Assessment Test (FCAT). We focus only on math and reading scores for this paper because those tests are given to all students in grades 3-10. The FCAT is given in writing and science to a subset of grades, but we do not use these data. We standardize all scores to have a mean of zero and a standard deviation of one within each grade-year combination. Third, we link students to teachers using a database that contains the courses taken by each student and the courses taught by each teacher. A unique classroom identifier also allows us to generate classroom measures, such as percent black and Hispanic, percent of students eligible for free or reduced price lunch, and average prior achievement, all of which we use as controls in the value-added models. We use this dataset to answer research questions 1 and 2. To answer research question 3, we append two teacher characteristics to the dataset: Spanish fluency and whether a teacher has a bilingual certification.[2] We obtain these teacher characteristics from teacher surveys that we administered in M-DCPS in 2010 and 2011.

Table 1 gives the proportion of EL students in M-DCPS during our sample period, as well as shows how this proportion varies by grade. For Tables 1 and 2, which describe the sample, we use the original definition of EL from the administrative dataset to facilitate easier

[2] Teachers are considered to have a bilingual certification if they indicate on the survey that "Foreign Language/Bilingual" is an area in which they are certified to teach. In Florida, the certification subjects that fit under this category are World Languages or English for Speakers of Other Languages (ESOL). Teachers who teach ELs must have or be in the process of obtaining ESOL certification/training. The courses that are a part of the ESOL endorsement are Testing and Evaluation of ESOL, Cross Cultural Communication and Understanding, Methods of Teaching ESOL, ESOL Curriculum and Material Development, and Applied Linguistics (see http://bilingual.dadeschools.net/BEWL/briefings_memos.asp and http://bilingual.dadeschools.net/BEWL/META/Info.asp for additional information).

comparability with students from other districts and states. Between the 2003-04 and the 2009-10 school years, the proportion of ELs remained fairly constant around .095, with a slight uptick in 2010-11 to 0.125. Grade 3 consistently has the highest proportion of ELs with a general though inconsistent decline across the higher grades. In Florida, English learners are exempt from testing if they have been enrolled in school in the United States for less than 12 month. As expected from national trends in EL performance on standardized tests, a substantial gap in test scores can be seen between ELs and non-ELs.

Table 2 describes our sample at the student-, class-, and school-level, overall, for EL students and for non-EL students. Not surprisingly, ELs are more likely than non-ELs to be Hispanic and less likely to be black or white. Further, a higher percentage of ELs (80 percent) are eligible to receive free or reduced-price lunch compared to non-ELs (62 percent). Descriptive statistics at the class level also provide a picture of the students and teachers in M-DCPS. Over the span of our study, 62 percent of students in the average student's class were Hispanic and ten percent were EL. As for teachers, on average 41 percent were fluent in Spanish and five percent had a bilingual certification. EL students attend classes with a high proportion of Hispanic and poor students, but a lower proportion of special education students than non-EL students, on average.

Methods

In this study, we create separate value-added measures of teacher effectiveness for each teacher's impact on EL and non-EL students. We then use these separate measures to better understand teacher effectiveness for ELs by addressing the following research questions:

1) How much does teacher effectiveness vary across classrooms for EL and non-EL students?

2) Are teachers equally effective with ELs and non-ELs?
3) Can measured teacher characteristics help explain these differences in value-added? In particular, are teachers who have bilingual certification or are fluent in Spanish differentially more effective with English learners?

Estimating Value-Added.

The study relies on value-added measures of teacher effectiveness. As discussed above, these measures are common in both research and practice, though there is no consensus on the best method for estimating value-added. Regardless of the particular estimation technique used, the goal of value-added measures is to isolate the effects of the classroom teacher from the effects of student background characteristics, peer effects, and school effects.

We calculate value-added estimates in the form of coefficients on teacher fixed effects used to predict student test score gains. Our approach constrains the estimates to sum to zero, which means teachers are compared to the average for a specified peer group rather than to an arbitrarily omitted teacher. For all of our teacher fixed effects models, we calculate value-added for ELs and non-ELs separately in order to compare the estimates. Further, we only run models for teachers who have ten or more students in either category across the seven years of data that we use for the analyses to ensure the estimates are based on a sufficient number of observations.[3] We do not estimate transitory teacher effects (i.e. different estimates for teachers in each year) because small EL sample sizes make such estimates unstable.

Specifically, we estimate a teacher fixed-effects model, as described by Equation 1, that predicts the test scores in year t for student i in grade g with teacher j in school s as a function of the test score in year $t-1$, student ($\mathbf{X}_{ijst}$), classroom ($\mathbf{C}_{jt}$), and school ($\mathbf{S}_{st}$) characteristics (for a

[3] We also did not make any limitation on the number of classes taught by a teacher because the number of teacher-class-year observations made up by teachers with only one class of data across all years is small (1% in both math and reading).

detailed list, see Appendix 2).[4] Such controls are included in order to mitigate bias that might result from the assignment of teachers to students with similar prior test scores but different propensities to learn during the course of the year. In addition to teacher fixed effects (δ_j), we also include year (γ_t) and grade (α_g) fixed effects to control for unobservable differences in test score gains due to variance from year to year (such as a district-wide policy change) and differences in test-score gains that occur from one grade to the next (such as a more difficult assessment being used). More broadly, these fixed effects control for differences in test score distributions that naturally occur from year to year and grade to grade. Finally, we control for a vector of prior year student test scores ($\mathbf{A}_{igjs(t-1)}$) in both math and reading.[5] For simplicity, we omit subscripts for academic subject, but we estimate the model separately for math and reading. Within each subject, we estimate Equation 1 twice: once using only EL students and once using non-EL students.

$$A_{igjst} = \mathbf{A}_{igjs(t-1)}\,\beta_1 + \mathbf{X}_{it}\,\beta_2 + \mathbf{C}_{jt}\,\beta_3 + \mathbf{S}_{st}\,\beta_4 + \delta_j + \gamma_t + \alpha_g + \varepsilon_{igjst} \qquad (1)$$

After estimating the models separately for ELs and non-ELs in both subjects, we use a Bayesian shrinkage procedure whereby we weight the mean of teacher value added more heavily as the standard error for a teacher's individual value added estimate increases (see Appendix 1 for a description of the method).

In what follows, we use teacher fixed-effects estimates to compare relative teacher efficacy with ELs and non-ELs. We then use regression models predicting test scores and controlling for lagged test scores to investigate which teacher characteristics are associated with higher gains. For each research question, we provide a general description of the models and analytical approaches used below.

[4] We look at teacher effects across schools, so school fixed effects are not included in the model.
[5] While a student fixed effect model may theoretically do a better job of isolating the teacher effect, we do not use such a model because of its imprecision and potential bias (Kane and Staiger, 2008).

Research Question 1: How much does teacher effectiveness vary across classrooms for EL and non-EL students?

Research question 1 asks whether teachers are more important for the achievement gains of ELs or non-ELs, which helps inform discussions about whether ELs are differentially affected by educational inputs. For example, if the variance in the teacher effects is much larger for ELs then the consequence of having a teacher in the top quartile of effectiveness would be much more beneficial for an EL student than for another student. Similarly, having a teacher in the bottom quartile of effectiveness would, in this hypothetical case, be much more detrimental for ELs.

To answer this question, we compare the variances of the teacher fixed-effects estimates in math and reading for teachers of ELs and non-ELs. We report this comparison for the "true" value-added estimates, which back out measurement error.[6] The true estimate is derived by taking the mean of the square of all the standard errors for individual teacher fixed-effects estimates, then subtracting that mean from the variance of the fixed effects. This approach removes the proportion of a teacher's value added that is due to measurement error.

To determine whether the variances for ELs and non-ELs are significantly different from each other, we use a bootstrapping approach. If we were interested in testing the equality of variances for the original value-added estimates, we could simply use a Levene test. However, we cannot use this same test for the true estimates because we back out the mean standard error of the estimate from the variance of the entire sample and therefore no longer have a distribution on which the Levene test can be performed. As a baseline to which we can compare the true variances, we compare our estimates to true estimates for groups of randomly generated ELs and

[6] Though not reported, we estimated the variances of the raw and shrunk estimates of the fixed effects. As should be expected, the variances of the raw scores are greater than the true variances and the variances of the shrunk scores are smaller than the true variances.

non-ELs. We generate these "random" ELs by determining what percent of a given teacher's class each year is composed of ELs, and then randomly assigning students to EL status in the same proportion. We repeat this process 75 times so that we have a distribution of variances for random ELs and non-ELs. We can then see where the variances for ELs and non-ELs falls on the sampling distribution and where the difference in variance between the two groups falls on the sampling distribution for the random differences.

Research Question 2: Are teachers equally effective with ELs and non-ELs?

We address this question in two ways. First, we correlate value-added gains (including both original Pearson correlations and the same estimates corrected for attenuation) for teachers of ELs and non-ELs separately for math and reading. As we do for the prior question, we also present correlations using the administrative definition of EL. Second, we cross-tabulate EL and non-EL value-added estimates by quintile. In essence, this combination provides a parametric (correlations) and non-parametric (cross-tabulations) method of examining the association between value-added estimates for teachers with their ELs and non-ELs. Though the non-parametric approach does not correct for attenuation, it closely resembles the sort of approach to categorizing teachers by effectiveness used by school systems, including those using estimates to make decisions about teacher promotion, retention, and remediation.[7] As a way to combine these two approaches, we estimate Spearman correlations—a non-parametric measure of association—across the entire sample and by quintile of teacher effectiveness with ELs and non-ELs. We find

[7] While one might worry that these quintile comparisons also underestimate differences that would occur in practice with less than seven years of data, the average years of value-added data for teachers with non-ELs is 3.5. With ELs, the average is 2.7 years. While the differences reported might be underestimated compared to estimates of value added that use only two years of data, that underestimation would be much less pronounced for school systems that pool estimates across more than two years.

Spearman correlations that are similar in magnitude to Pearson correlations, therefore we only report the latter.[8]

In keeping with the strategy we use for question one to determine the significance of differences in VA estimates, we compare correlations between value-added estimates for ELs and non-ELs to the same correlations from our randomly generated sets of ELs and non-ELs. If the correlations are similar, then the differences between a teacher's estimated value added with ELs and his or her estimated value added with non-ELs can be attributed largely to measurement error. However, if the correlation between teachers' value-added with ELs and non-ELs is lower than between randomly generated student groups, then some teachers are likely differentially effective with ELs. We supplement the analysis with random effects estimates which, though more parameterized, allow for the direct measure of the correlation between the two groups accounting for measurement error.

Research Question 3: Can measured teacher characteristics help explain differences in teacher effectiveness?

To better understand the source of differences in value-added estimates, we regress teacher characteristics on student test performance. Specifically, our models include covariates for a teacher's Spanish fluency and attainment of a bilingual certification. We do not include these or other teacher characteristics in Equation 1 because we are interested in the teacher effect, not the effect of teachers relative to peers who share certain characteristics. In this part of the analyses, however, we are interested in whether teachers with certain characteristics are more effective.

[8] Just as the overall Spearman correlations do not appear to differ more than expected from the overall Pearson correlations, the Spearman correlations do not appear to differ much across quintiles. These non-parametric correlations by quintile are, however, much lower than the overall correlations, due to reduced sample size.

Our student achievement models include largely the same controls used in our value-added models. We also include the teacher characteristic of interest and an interaction between EL status and that particular characteristic.[9] In the base model, we include grade and year fixed effects as controls. The specification is detailed in Equation 2, which includes the characteristic of interest, κ, and its interaction with EL status.

$$A_{igjst} = \mathbf{A}_{igjs(t-1)}\,\beta_1 + \mathbf{X}_{it}\,\beta_2 + \mathbf{C}_{jt}\,\beta_3 + \mathbf{S}_{st}\,\beta_4 + \beta_5\kappa_j + \beta_6\kappa_j * EL + \gamma_t + \alpha_g + \varepsilon_{igjst} \quad (2)$$

In order to account for the non-random sorting of teachers into schools that may be associated with the characteristic of interest, we run another specification similar to Equation 2 that includes a school fixed effect. These fixed effects allow us to compare how student achievement varies across teachers with different characteristics within the same school. Lastly, we run a third model with teacher fixed effects, which mitigates the potential bias of non-random assignment of students to teachers. This last specification allows us to compare the academic performance of EL and non-EL students within a teacher's classroom to investigate whether a teacher with specific attributes is more effective with one group. Note that in the teacher fixed effect model, the teacher characteristic of interest is omitted because it is absorbed by the teacher fixed effect.

Ideally we could randomly assign students to teachers so that we would not be concerned with omitted variables bias. However this approach was not viable. The methods described above, however, take a large step in accounting for potential omitted variables. First, rich prior controls in our regression models correct for selection based on measured characteristics such as prior test scores and absences. Second, the school fixed-effects models compare teachers within the same school and thus remove potential biases from the sorting of teachers and students to schools based on unmeasured characteristics. Finally, the teacher fixed-effects model adjusts for

[9] We do not find differential returns to experience, and thus do not include teacher experience in equation 2.

the systematic sorting of students to teachers even within the same school and even on unmeasured characteristics. The remaining potential source of bias is *differential* sorting of ELs and non-ELs on *unmeasured* characteristics to the same teacher. While it is feasible that, for example, EL students who would be more likely to make gains because of unmeasured characteristics are systematically sorted to teachers with non-ELs who are less likely to make gains because of unmeasured characteristics, this concern is unlikely to be as great as the two issues addressed above by school and teacher effects. The results of these analyses will not be the final word on the relationship between teacher characteristics and differential effectiveness with ELs but they provide initial examination of likely hypotheses for these differential effects.

Findings

Research Question 1: How much does teacher effectiveness vary across classrooms for EL and non-EL students?

Table 3 shows the standard deviations of each different set of value-added measures we estimate, i.e. each combination of math or reading and of EL or non-EL. As discussed in the methods section, for each set, we report the standard deviation of the "true" value-added estimates from which we have subtracted measurement error. Our findings dovetail with those produced in other value-added research (Hanushek & Rivkin, 2010). Specifically, like Hanushek and Rivkin (2010) who present the standard errors of their shrunk fixed-effects estimates, we find a shrunk standard deviation in math value-added of .10 (not reported).

The "true" values are our best estimates of the actual variance of value-added. While, in most cases, the estimates of variance are greater for teachers of non-ELs than ELs, the differences are small and similar to what we would expect given a random draw from similar

populations with equal variances.[10] For example, when math achievement is used as the outcome in Equation 1, the standard deviation of the true teacher effects is approximately .15 for ELs and .17 for non-ELs for a difference of .02. These differences are even smaller in magnitude when using the administrative definition of EL. For all students, the difference in standard deviations between ELs and non-ELs is -.020 in math and -.003 in reading.

To assess whether the differences in standard deviations are statistically significant for the true estimates, the last column of Table 3 shows the standardized difference based on 75 runs in which we randomly generated ELs and estimated their true standard deviations. We find that, except in high school math, which may be an anomaly since it is just one test of eight, there is no significant difference in the true variance in value-added of teacher effects for non-ELs and ELs. In analyses not presented, we compare the variance estimates obtained in our fixed effect specification to those obtained using a random coefficients model in which the true variance for the two groups is directly estimated from the model. When run for ELs and non-ELs separately, the difference in the variance estimates is similar to that of the fixed effects model for math, and slightly larger for reading.[11]

The results provide evidence that the variances are similar for ELs and non-ELs, and that observed differences are likely due to measurement error. We also check that this finding is robust to whether we estimate the distributions only using teachers with estimates for both types of students or if we use all teachers with available data, and find no observable differences in the results.

Research Question 2: Are teachers equally effective with ELs and non-ELs?

[10] We also estimate the variances in a random effects framework and found similar results: for ELs, the standard deviations in reading and math were .103 and .156, respectively. For non-ELs, the standard deviations in reading and math were .110 and .173, respectively.

[11] Similarly, the standard deviations of the teacher effect are approximately the same for teaching EL students and non-EL students when using the administrative definition of English learner status.

All of our models produce high correlations between value-added for ELs and non-ELs, though not as high as for randomly generated groups of students. Teachers who are good with ELs tend to be good with non-ELs and vice-versa, though some teachers are somewhat better with one group than the other.

Tables 4 and 5 use value-added estimates from Equation 1 in math and reading, respectively, to show a transition matrix of teachers' value-added for ELs and non-ELs by quintile. First looking at the matrix for math, 59 percent of the teachers in the top quintile of value-added for non-ELs are also in the top quintile of value-added for ELs. Of those teachers in the bottom quintile for non-ELs, 50 percent are in the bottom quintile for ELs. These results suggest there is significant overlap in teachers who are effective with ELs and non-ELs. Similarly, less than four percent of teachers are either in both the top quintile for non-ELs and in the bottom quintile for ELs or in both the bottom quintile for non-ELs and in the top quintile for ELs. Very few teachers have high value-added for one group and low value-added for the other group.

The overlap for reading is not as great as for math, but there is still substantial overlap. Forty two percent of teachers in the top quintile for non-ELs are in the top quintile for ELs and 35 percent of teachers in the bottom quintile for non-ELs are in the bottom quintile for ELs. Only seven percent of teachers who are in the top quintile for non-ELs are also in the bottom quintile for ELs and, again, only seven percent of teachers who are in the bottom quintile for non-ELs are also in the top quintile for ELs.

Table 6 presents the correlations between value-added for ELs and non-ELs by school level. Part of reason for the lack of overlap evident in the cross-tabulations comes from measurement error. In order to address this issue, Table 6 includes attenuation-corrected

correlations, which tend to be much higher than the uncorrected Pearson correlations. We find a correlation of 0.89 for Math and 0.80 for Reading.[12] The attenuation-corrected correlations are also high—though not as high—when using the administrative definition of ELs, roughly .65 in math and reading. In keeping with Table 4, the correlation for math is higher than for reading in elementary and middle school, though not in high school. While the dis-attenuated correlations are meaningfully higher than those without the correction, the correlations are still imperfect (i.e. less than 1.0). Our analysis of randomly generated groups of students confirms this conclusion. When we randomly generated a group of ELs in the same proportion as is actually in a teacher's classroom, the correlation between value-added for ELs and non-ELs is generally higher than what we get with actual ELs and non-ELs, providing evidence that some teachers are somewhat better with one group than the other. The last column of Table 5 shows how great the observed correlation is relative to the sampling distribution of correlations from random draws. We want to know whether we could have obtained the correlations we did just from drawing two groups of similar students instead of one group of ELs and one of non-ELs. In fact, across all school levels in math and reading, we see that we would have been unlikely to draw two correlations as low as we did. For example, while we find a correlation of 0.61 between EL and non-EL value-added, the average correlation from random draws is 0.67 with a standard deviation of 0.01. Thus, the difference between the actual and the random is greater than three standard deviations. In reading, the difference is over two standard deviations of the sampling distribution difference.[13]

[12] We also estimate these correlations in a random effects framework and find similar results, including for the attenuation-corrected correlations: .83 in reading and .86 in math.

[13] Using the administrative definition of English learner status we find somewhat lower but still robust correlations between teacher value-added to test performance for their ELs and for their non-ELs (approximately 0.7 for both math and ELA, disattenuated).

These findings provide evidence that our imperfect correlations for ELs and non-ELs are not due entirely to measurement error. If correlations between real ELs and non-ELs were largely the result of measurement error, then they would be closer to those generated for random groups. The lower correlation in value-added between real ELs and non-ELs compared to randomly generated ELs and non-ELs suggests that there are likelier to be actual differences in value-added by group, though the differences are not great.[14]

Research Question 3: Can measured teacher characteristics help explain differences in teacher effectiveness?

Tables 7 and 8 show the results of student-level regression analyses that predict student achievement (in math and reading, respectively) as a function of teacher characteristics described in Equation 2. The tables give results from models that include: (1) no fixed effects, (2) school fixed effects, and (3) teacher fixed effects. The coefficients presented are the regression coefficients from the interaction of EL with the relevant teacher characteristic. Because student test scores are the outcomes, such a coefficient tells us what the achievement gap is between ELs and non-ELs when they have a teacher with a particular characteristic. For example, at the elementary level, ELs experience a .10 standard deviation increase in math achievement over their non-EL counterparts when they have a teacher who is fluent in Spanish, and a .18 standard deviation gain with a bilingually certified teacher.

In both math and reading, we see that all but one of the estimates of the teacher characteristic interacted with EL in the table that are significantly different from zero are positive, indicating that teachers who are fluent in Spanish or have a bilingual certification are

[14] We compare the correlations of the teacher effects obtained in the fixed effect specification with those obtained using a random coefficients model in which the correlation is estimated directly from the model. The results are highly significant correlations of .84 and .74 in math and reading, respectively, corroborating the findings from our fixed effect specification.

more effective with ELs relative to non-ELs. The effect of Spanish fluency is less pronounced in reading than in math, and the opposite is true for bilingual certification. The coefficients for bilingual certification at the elementary level are especially large from a practical standpoint (over one-tenth of a standard deviation in all cases) and are significant in all model specifications across math and reading. In general, these results hold up for models including a teacher fixed effect, which controls for the non-random sorting of students to teachers.

Discussion & Conclusions

This study asks whether teachers who are effective at teaching English learners are the same teachers as those who are effective at teaching English-proficient students. We first find little discernible difference in the importance of teachers for the achievement gains of ELs and non-ELs. That is, the variation in teacher effectiveness is generally as great for ELs as it is for non-ELs. We also find that teachers who are effective with one group also tend to be effective with the other group. This said, some teachers are somewhat more effective with one group or the other. The two teacher characteristics that we test – language proficiency in the students' first language and bilingual certification – both predict differential positive effectiveness with English learners.

The implications of the results are two-fold. First, if a goal is to improve outcomes for English learners and a choice is to assign teachers who are relatively more effective on average than other teachers or to assign teachers who appear to be relatively more effective with English learners than with English proficient students, then the first choice is likely to lead to better outcomes for English learners. That is, finding a better teacher for English learners is at least as much if not more a question of finding an effective teacher, as it is a question of finding a teacher

who specializes in English learners. The differential effectiveness of teachers with English learners is a relatively small part of what makes a teacher good with English learners.

The second implication of the results is that even though the differential effectiveness of teachers with English learners does not explain a lot of what makes a teacher good with English learners, we find suggestive evidence that there are specific skills that can boost teachers' effectiveness with English learners. In particular, though not surprising, speaking the student's first language appears important, as does bilingual certification.

Finally, the findings raise some questions for teacher evaluation. The correlation of the true value-added for ELs and non-ELs is strong but not perfect. As a result, teachers who would be classified in one way if rated only on their effectiveness with English learners could be classified in another way if rated only on their effectiveness with other students or based on their average effectiveness score with all students. For illustrative purposes, we estimate the extent of this misclassification, given the numbers in this study and assuming that teachers are classified into four equal groups based on their value-added estimate. First compare teachers' value-added scores for ELs to their average value-added scores with all their students. Approximately 40 percent of teachers would be differently classified using these different value-added measures, though most of the misclassification would be between contiguous groups – not, for example, from the lowest group in one classification to the top half using the other value-added measure. For example, using math value-added, approximately six percent of teachers who are actually in the least-effective group with their EL students would be classified in the second to lowest quartile on the basis of the observed value-added with all of their students, and a little less than one percent of these teachers would be classified as being in the top 50 percent of teachers. Similarly, of teachers who are truly in the second lowest group with the ELs, about five percent

would be classified as in the lowest group using observed value-added with all of their students, while about seven percent of teachers in this second-to-bottom group would be classified as being in the top half of the distribution. Misclassification is even worse comparing value-added with ELs to value-added with non-ELs with approximately 55 percent of teachers misclassified, though again, the misclassification is mostly between contiguous groups. Moreover, because value-added measures are less precise with ELs (or with other specific student groups) than they are with all students, how an evaluation system adjusts for measurement error can affect teachers' value-added estimates differentially for estimates based only on ELs than for the estimates based on larger samples. The imprecision in the unadjusted EL value-added scores means that teachers will be more likely to receive extreme scores than they would be with measures for non-ELs. Conversely, if the evaluation system uses shrunk scores, the smaller variance for value-added scores based on ELs will mean that teachers will be less extreme. The value-added estimates could be standardized to maintain comparability, but this is an extra step in the process.[15] In any case, the clearly large classification differences and issues of measurement error point to the drawbacks of relying heavily on value-added groupings for evaluations.

Like all studies, this one is clearly imperfect. A number of issues stand out. First, the study was conducted in Miami-Dade County Public schools. The English learner population in this district differs from that in some large districts in that the vast majority is Spanish speaking. This homogeneity has implications for instruction in comparison to districts with smaller and more varied English learner populations. Similarly, many English proficient students also speak Spanish as do many adults in schools. These are some of many characteristics that might make teaching and teaching effectiveness different in MDCPS than elsewhere. Second, we have only

[15] We thank an unidentified reviewer of this paper for this insight and analysis.

lightly touched on characteristics of teachers that might be associated with differentially more effective teaching for English learners. The contribution of this paper is that it shows that this differential effect is only a relatively small part of the total effectiveness of teachers with English learners. Definitively showing which teacher characteristics drive differential effectiveness is beyond the scope of this paper because it requires more focus on the nuances of teaching and learning and developing a strategy for estimating causal effects than this paper can provide. Finally, the research literature on value-added modeling is very much in development. While there is a strong research base to support the approaches we have taken here, it was beyond the scope of a single paper to assess the implications of all model attributes for our findings. As our understanding of modeling improves, the best choice for modeling our research questions may also change. Further analysis could expand the value-added models as well as expand the geographic scope of the analyses and the causal analysis of factors affecting school and teacher value-added with English learners.

References

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

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1), 1-28.

August, D., & Pease-Alvarez, L. (1996). Attributes of Effective Programs and Classrooms Serving English Language Learners.

August, D., & Shanahan, T. (2007). *Developing Reading and Writing in Second Language Learners: Lessons from the Report of the National Literacy Panel on Language-Minority Children and Youth*. Taylor & Francis.

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

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6), 673–682.

Dee, Thomas S. (2005). A Teacher like Me: Does Race, Ethnicity, or Gender Matter? *The American Economic Review*, 95(2), 158–165.

Dee, Thomas S. (2007). "Teachers and the Gender Gaps in Student Achievement." *Journal of Human Resources*, 42(3): 528-554.

Feng, L. (2010). Hire today, gone tomorrow: New teacher classroom assignments and teacher mobility. *Education Finance and Policy*, *5*(3), 278-316.

Fry, R. (2007). *How far behind in math and reading are English language learners?* Washington, DC: Pew Hispanic Center.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*, 100(2), 267–271.

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

Hill, H. C. (2007). Learning in the Teaching Workforce. *The Future of Children*, 17(1), 111–127.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, 42(2), 371–406.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (Working Paper #11463). National Bureau of Economic Research.

Kalogrides, D., Loeb, S., & Beteille, T. (Forthcoming). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*.

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

Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation* (Working Paper No. 14607). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14607

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. *Policy and practice brief prepared for the Bill and Melinda Gates Foundation. Posted: March*.

Kao, G., & Thompson, J. S. (2003). Racial and Ethnic Stratification in Educational Achievement and Attainment. *Annual Review of Sociology*, 29, 417–442.

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring Student-Teacher Interactions in Longitudinal Achievement Data. *Education Finance and Policy*, 4(4), 439–467.

Loeb, S., & Candelaria, C. A. (2012). How Stable Are Value-Added Estimates across Years, Subjects and Student Groups? What We Know Series: Value-Added Methods and Applications. Knowledge Brief 3. *Carnegie Foundation for the Advancement of Teaching*.

Loeb, S., Kalogrides, D., & Béteille, T. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy*, 7(3), 269-304.

Master, B., Loeb, S., Whitney, C., & Wyckoff, J. (2012). Different Skills? Identifying Differentially Effective Teachers of English Language Learners. *Manuscript submitted for publication.* Retrieved from http://cepa.stanford.edu/sites/default/files/ELL%20Teacher%20Effects%20March%202012.pdf

McCaffrey, D. F. (2012). Do Value-Added Methods Level the Playing Field for Teachers? What We Know Series: Value-Added Methods and Applications. Knowledge Brief 2. *Carnegie Foundation for the Advancement of Teaching*.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis*, 26(3), 237–257.

Parrish, T. B., Merickel, A., Perez, M., Linquanti, R., Socias, M., Spain, A., ... & Delancey, D. (2006). Effects of the Implementation of Proposition 227 on the Education of English Learners, K-12: Findings from a Five-Year Evaluation. Final Report for AB 56 and AB 1116. American Institutes for Research and WestEd.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White Achievement Gap in Math and Reading in the Elementary Grades. *American Educational Research Journal*, 46(3), 853–891.

Reardon, S.F., & Raudenbush, S.W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, 4(4), 492-519.

Rice, J. K. (2003). Teacher Quality: Understanding the Effectiveness of Teacher Attributes. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED480858

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one?. *Education Finance and Policy*, *6*(1), 43-74.

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

Rothstein, J. (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, 4(4), 537–571.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, 29(1), 103–116.

Rumberger, R., & Gandara, P. (2004). Seeking Equity in the Education of California's English Learners. *Teachers College Record*, 106(10), 2032–2056.

Slavin, R. E., & Cheung, A. (2005). A Synthesis of Research on Language of Reading Instruction for English Language Learners. *Review of Educational Research*, 75(2), 247–284.

Solomon, M., Lalas, J., & Franklin, C. (2006). Making instructional adaptations for English learners in the mainstream classroom: Is it good enough. *Multicultural Education*, 13(3), 42-45.

Téllez, K., & Waxman, H. C. (2006). Preparing quality teachers for English language learners: An overview of the critical issues. *Preparing quality teachers for English language learners: Research, policies, and practices*, 1-22.

Wayne, A. J., & Youngs, P. (2003). Teacher Characteristics and Student Achievement Gains: A Review. *Review of Educational Research*, 73(1), 89–122.

Williams, T., Hakuta, K., Haertel, E., Kirst, M., Perry, M., Oregon, I., Brazil, N., et al. (2007). *Similar English Learner Students, Different Results: Why Do Some Schools Do Better? A follow-up analysis based on a large-scale survey of California elementary schools serving low-income and EL students*. Mountain View, CA: EdSource. Retrieved from http://www.edsource.org/assets/files/SimELreportcomplete.pdf

Table 1

Proportion of students who were ELs and standardized test scores in MDCPS, by year

	Year							
	2003-04	**2004-05**	**2005-06**	**2006-07**	**2007-08**	**2008-09**	**2009-10**	**2010-11**
Percentage of ELs	0.102	0.096	0.090	0.089	0.091	0.091	0.105	0.125
Percentage of ELs by grade								
Grade 3	0.149	0.139	0.130	0.140	0.142	0.152	0.171	0.237
Grade 4	0.095	0.089	0.086	0.082	0.084	0.091	0.132	0.155
Grade 5	0.091	0.090	0.075	0.078	0.074	0.079	0.090	0.125
Grade 6	0.082	0.081	0.081	0.070	0.073	0.070	0.081	0.095
Grade 7	0.095	0.087	0.082	0.086	0.078	0.075	0.085	0.095
Grade 8	0.096	0.093	0.083	0.081	0.090	0.075	0.084	0.090
Grade 9	0.101	0.092	0.086	0.085	0.089	0.086	0.088	0.101
Grade 10	0.103	0.098	0.093	0.090	0.091	0.093	0.103	0.100
Standardized math test scores	-0.549	-0.591	-0.605	-0.630	-0.618	-0.648	-0.648	-0.605
Standardized reading scores	-0.829	-0.898	-0.942	-0.918	-0.945	-0.938	-0.916	-0.891

Table 2

Student, Class, and School Characteristics, by EL status

	Overall		EL		Non-EL	
Student	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
White	0.10	-	0.02	-	0.10	-
Black	0.26	-	0.12	-	0.27	-
Hispanic	0.62	-	0.85	-	0.60	-
Free or reduced price lunch	0.64	-	0.80	-	0.62	-
Special education	0.10	-	0.07	-	0.11	-
Ever EL	0.10	-	1.00	-	0.00	-
Math Score	0.04	0.98	-0.61	1.08	0.11	0.94
Reading Score	0.04	0.98	-0.91	1.00	0.15	0.92
Class	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
White	0.09	0.13	0.02	0.10	0.10	0.14
Black	0.26	0.33	0.13	0.29	0.28	0.33
Hispanic	0.62	0.32	0.84	0.31	0.59	0.32
Free or reduced price lunch	0.64	0.26	0.79	0.25	0.63	0.26
Special education	0.10	0.23	0.07	0.22	0.10	0.24
Ever EL	0.10	0.25	1.00	0.00	0.00	0.00
Math Score	-0.04	0.75	-0.94	0.75	0.06	0.70
Reading Score	-0.06	0.78	-1.30	0.69	0.07	0.70
Spanish fluent teacher	0.41	-	0.63	-	0.37	-
Bilingual certified teacher	0.05	-	0.19	-	0.02	-
School	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
White	0.09	0.11	0.02	0.06	0.10	0.11
Black	0.26	0.31	0.12	0.27	0.27	0.31
Hispanic	0.62	0.29	0.85	0.27	0.60	0.29
Free or reduced price lunch	0.64	0.22	0.80	0.18	0.62	0.23
Special education	0.10	0.06	0.07	0.07	0.10	0.06
Ever EL	0.10	0.07	1.00	0.00	0.00	0.00
Math Score	0.03	0.36	-0.61	0.39	0.10	0.36
Reading Score	0.03	0.36	-0.91	0.35	0.14	0.35

Note. Numbers in table represent averages at the student-year level.

Table 3

Comparing "true" value-added standard deviations by School Level for ELs and Randomly Generated ELs

	EL	Non-EL	Difference	Random Difference[a]	SDs Apart
All Grades[b]					
Math	0.153	0.171	0.018	0.025 (0.012)	-0.609
Reading	0.11	0.127	0.017	0.015 (0.011)	0.237
Elementary					
Math	0.23	0.212	0.015	0.008 (0.011)	0.587
Reading	0.144	0.142	0.002	0.009 (0.011)	-1.000
Middle					
Math	0.141	0.163	0.022	0.031 (0.011)	-0.785
Reading	0.083	0.094	0.011	0.008 (0.016)	0.165
High					
Math	0.149	0.128	-0.021	0.020 (0.013)	-3.182*
Reading	0.104	0.115	0.011	0.004 (0.009)	0.739

a. Randomly generated EL mean and standard deviations are based on 75 runs.
b. The true differences between ELs and non-ELs using the administrative definition of EL are -.020 in math and -.003 in reading.

Table 4

Correlations by quintile of value-added for teachers of ELs and Non-ELs (math)

EL VA	Non-EL VA					
	1	**2**	**3**	**4**	**5**	**Total**
1	204	132	80	36	15	467
	43.68	28.27	17.13	7.71	3.21	100
	49.88	25.14	14.79	7.14	3.69	19.57
2	93	170	124	71	22	480
	19.38	35.42	25.83	14.79	4.58	100
	22.74	32.38	22.92	14.09	5.41	20.12
3	62	121	149	104	48	484
	12.81	25	30.79	21.49	9.92	100
	15.16	23.05	27.54	20.63	11.79	20.28
4	35	79	133	152	81	480
	7.29	16.46	27.71	31.67	16.88	100
	8.56	15.05	24.58	30.16	19.9	20.12
5	15	23	55	141	241	475
	3.16	4.84	11.58	29.68	50.74	100
	3.67	4.38	10.17	27.98	59.21	19.91
Total	409	525	541	504	407	2,386
	17.14	22	22.67	21.12	17.06	100
	100	100	100	100	100	100

Note. Overall correlation is .6138.

Table 5

Correlations by quintile of value-added for teachers of ELs and Non-ELs (reading)

EL VA	Non-EL VA					
	1	**2**	**3**	**4**	**5**	**Total**
1	150	94	96	41	28	409
	36.67	22.98	23.47	10.02	6.85	100
	34.72	20.35	18.79	8.56	7.18	17.99
2	124	121	114	74	37	470
	26.38	25.74	24.26	15.74	7.87	100
	28.7	26.19	22.31	15.45	9.49	20.67
3	82	114	114	106	59	475
	17.26	24	24	22.32	12.42	100
	18.98	24.68	22.31	22.13	15.13	20.89
4	46	83	110	135	103	477
	9.64	17.4	23.06	28.3	21.59	100
	10.65	17.97	21.53	28.18	26.41	20.98
5	30	50	77	123	163	443
	6.77	11.29	17.38	27.77	36.79	100
	6.94	10.82	15.07	25.68	41.79	19.48
Total	432	462	511	479	390	2,274
	19	20.32	22.47	21.06	17.15	100
	100	100	100	100	100	100

Note. Overall correlation is .4384.

Table 6

Comparing correlations of teacher value-added scores for real versus randomly generated ELs

Correlations between ELs and Non-ELs for real and randomly generated ELs						
	EL Disattenuated/Original		Randomly generated EL[a] mean (SD)		Standardized Difference	
	Math	**Reading**	**Math**	**Reading**	**Math**	**Reading**
All[b]	.89/.61	.80/.44	.67 (.01)	.49 (.02)	-3.65	-2.16
Elementary	.97/.67	.78/.45	.71 (.03)	.49 (.03)	-1.15	-1.42
Middle	.89/.65	.75/.39	.66 (.02)	.45 (.03)	-0.55	-2.49
High	.59/.42	.80/.44	.48 (.03)	.48 (.03)	-2.22	-1.14

a. Randomly generated EL mean and standard deviations are based on 75 runs.
b. These correlations for the administrative definition of EL are .65/.44 in math and .65/.36 in reading.

Table 7

Results of regressions using teacher characteristics to predict test score gains (math)

Math	EL versus Non-EL Achievement Gap		
	No fixed effects	**School fixed effects**	**Teacher fixed effects**
All Levels			
Spanish Fluency * EL	0.055***	0.048***	0.044***
	(0.016)	(0.013)	(0.013)
Bilingual Certification * EL	0.039	0.020	0.039
	(0.054)	(0.056)	(0.055)
Elementary			
Spanish Fluency * EL	0.102***	0.107***	0.069**
	(0.029)	(0.025)	(0.021)
Bilingual Certification * EL	0.182**	0.176*	0.184***
	(0.067)	(0.069)	(0.036)
Middle			
Spanish Fluency * EL	0.033	0.022	0.038~
	(0.027)	(0.020)	(0.022)
Bilingual Certification * EL	-0.077	-0.106*	-0.070
	(0.048)	(0.050)	(0.059)
High			
Spanish Fluency * EL	0.020	0.019	0.024
	(0.028)	(0.02)	(0.025)
Bilingual Certification * EL	0.293	0.250	0.769
	(0.251)	(0.246)	(0.768)

*Note. ~p<.1, *p<.05, **p<.01, ***p<.001. Models include controls for student prior performance and demographic characteristics, comparable classroom average characteristics, and individual year and grade fixed effects.*

Table 8

Results of regressions using teacher characteristics to predict test score gains (reading)

Reading	EL versus Non-EL Achievement Gap		
	No fixed effects	School fixed effects	Teacher fixed effects
All Levels			
Spanish Fluency * EL	0.013	0.012	0.013
	(0.014)	(0.014)	(0.012)
Bilingual Certification * EL	0.066**	0.067**	0.050*
	(0.025)	(0.023)	(0.021)
Elementary			
Spanish Fluency * EL	0.044~	0.041~	0.044~
	(0.024)	(0.023)	(0.022)
Bilingual Certification * EL	0.100*	0.100~	0.154***
	(0.05)	(0.052)	(0.039)
Middle			
Spanish Fluency * EL	0.003	0.007	-0.004
	(0.024)	(0.021)	(0.02)
Bilingual Certification * EL	0.058~	0.066*	0.046*
	(0.031)	(0.032)	(0.023)
High			
Spanish Fluency * EL	0.003	-0.002	-0.001
	(0.023)	(0.022)	(0.020)
Bilingual Certification * EL	0.068	0.058	-0.042
	(0.045)	(0.044)	(0.056)

*Note. ~p<.1, *p<.05, **p<.01, ***p<.001. Models include controls for student prior performance and demographic characteristics, comparable classroom average characteristics, and individual year and grade fixed effects.*

Appendix 1

Details on Bayesian Shrinkage

Our estimated teacher effect ($\hat{\delta}_j$) is the sum of a "true" teacher effect (δ_j) plus some measurement error[16]:

$$\hat{\delta}_j = \delta_j + \varepsilon_j. \quad (2)$$

The empirical Bayes estimate of a teacher's effect is a weighted average of their estimated fixed effect and the average fixed effect in the population where the weight, λ_j, is a function of the precision of each teacher's fixed effect and therefore varies by j. The less precise the estimate, the more we weight the mean. The more precise the estimate, the more we weight the estimate and the less we weight the mean. Similarly, the more variable the true score (holding the precision of the estimate constant) the less we weight the mean, and the less variable the true score, the more we weight the mean assuming the true score is probably close to the mean. The weight, λ_j, should give the proportion of the variance in what we observe that is due to the variance in the true score relative to the variance due to both the variance in the true score and precision of the estimate. This more efficient estimator of teacher quality is generated by:

$$E(\delta_j \mid \hat{\delta}_j) = (1 - \lambda_j)(\bar{\delta}) + (\lambda_j) * \hat{\delta}_j \quad (3)$$

$$\text{where } \lambda_j = \frac{(\sigma_\delta)^2}{(\sigma_{\varepsilon j})^2 + (\sigma_\delta)^2} \quad (4)$$

Thus, the term λ_j can be interpreted as the proportion of total variation in the teacher effects that is attributable to true differences between teachers. The terms in (4) are unknown so are estimated with sample analogs.

$$(\hat{\sigma}_{\epsilon j})^2 = var(\hat{\delta}_{\varepsilon j}) \quad (5)$$

which is the square of the standard error of the teacher fixed effects. The variance of the true fixed effect is determined by:

$$(\sigma_\delta)^2 = (\hat{\sigma}_\delta)^2 - mean(\hat{\sigma}_\varepsilon)^2 \quad (6)$$

where $(\hat{\sigma}_\delta)^2$ is the variance of the estimated teacher fixed effects (Gordon, Kane, & Staiger, 2006; Jacob & Lefgren, 2005).

[16] Here we make the classical errors in variables (CEV) assumption, assuming that measurement error is not associated with an unobserved explanatory variable.

Appendix 2

Covariates for value-added models

- Lagged achievement in math and reading
- Race
- Gender
- Free and reduced price lunch (FRPL) status
- Whether the student was retained
- Special education status
- Lagged absences
- Lagged suspensions
- Grade dummies
- Year dummies
- Classroom Race proportions
- Classroom Gender proportion
- Classroom FRPL proportion
- Classroom English Learner proportion
- Mean classroom lagged achievement
- Mean classroom lagged absences
- Mean classroom lagged suspensions
- School FRPL proportion
- School Race proportions
- Mean school lagged achievement
- School enrollment

Covariates for student-level models

- Lagged achievement in math and reading
- Race
- Gender
- Free and reduced price lunch status
- Special education status
- Lagged absences
- Lagged suspensions
- Interaction between special education status and English Learner status
- Teacher Spanish fluency or teacher bilingual certification
- Grade dummies
- Year dummies
- Classroom Race proportions
- Classroom Gender proportion
- Classroom FRPL proportion
- Classroom English Learner proportion
- Mean classroom lagged achievement
- Mean classroom lagged absences
- Mean classroom lagged suspensions
- School FRPL proportion
- School Race proportions
- Mean school lagged achievement
- School enrollment