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

Value-added models are being used with increasing frequency to evaluate educational policies and programs, as well as teachers 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 with ELs is predicted by a teacher's fluency in Spanish 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. For example, 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 fully English proficient (FEP) students. In particular, we ask four research questions: (1) How much do teachers explain differences in learning among ELs and is this different than among FEPs? (2) To what extent are teachers who are effective with ELs also effective with FEPs? (3) How much of the variation in teacher effectiveness occurs within versus between schools for ELs and FEPs? Finally, (4) can measurable 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 FEP students. We also find that, on average, teachers who are effective with FEP 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 FEPs.

Background

Research consistently shows that teachers can meaningfully affect student achievement. The work of Aaronson, Barrow, and Sander (2007) quantifies the effect: a one standard deviation improvement in teacher quality for one year raises student math scores by roughly two-fifths of the average gain made over the same period. In a similar vein, Sanders and Rivers (1996) examine the impact of having different sequences of teachers on academic achievement and show that students with a string of effective teachers have a significant advantage. Other studies on teacher effects agree that teachers matter a great deal for scholastic outcomes (Hanushek, 2011; Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005). Furthermore, the research provides evidence that the quality of a student's teacher can have long-term job market implications. Specifically, Chetty, Friedman, and Rockoff (2011) find that a one standard deviation improvement in teacher value added raises earnings by about one percent at age 28 and that replacing a teacher whose value added is in the bottom five percent with an average teacher would increase a student's lifetime income by more than $250,000.

Given the importance of teachers to student achievement, much research is devoted to determining how best to measure teacher quality. There are many possible measures of teacher effectiveness from student critiques (Ferguson, 2010), to informal principal evaluations (Jacob & Lefgren, 2008), to formal measures based on observational protocol (Danielson, 2007; Grossman, Loeb, Cohen, & Wyckoof, forthcoming; Pianta, Paro, & Hume, 2007). Measures of value-added to student achievement - the amount teachers increase the achievement test scores of their students over the course of the year – has 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, of being relatively low-cost to calculate for some teachers given the testing regimes already in place, and of 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, research suggests 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 most other measures of teacher effectiveness (McCaffrey, 2012; Rothstein, 2009). Nonetheless, value-added is far from a perfect measure of teacher quality (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders & Horn, 1994). It is highly dependent on the exam on which it is based (Kane & Staiger, 2012; Lockwood et. al., 2007; Papay, 2011), and it is imprecise, containing substantial measurement error (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). While, there is some evidence that value-added tends to

be correlated with better classroom practices as evaluated by other teacher quality measures, these relationships are not always strong (Grossman, Loeb, Cohen, & Wyckoff, forthcoming; Kane & Staiger, 2012). However, the desirable traits of value-added measures have led to their increasingly widespread use in research and practice. For example, value-added measures underlie major teacher policies in districts like New York City, states like Florida and Tennessee, and at the federal level, especially via criteria for states to receive funding under the Race to the Top program.

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 all students (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 generally effective, 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 FEP students. We choose ELs because they are a rapidly growing subgroup with unique educational challenges and, therefore, may benefit from EL-specific instructional strategies (Master, Loeb, Whitney, & Wyckoff, 2012). 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 FEP students (Fry, 2007). On average, at the end of high school, students who are still classified as English learners are almost five years behind FEP students on standardized tests (Parrish et al., 2006; Rumberger & Gandara, 2004). ELs prove less likely to progress through school than any other student subgroup (Kao & Thompson, 2003). First, second, and third generation Latino immigrants have high school dropout rates above twenty percent (Perreira, Harris, & Lee, 2006) and are less likely to get a post-secondary degree than any other racial/ethnic group in the U.S. (Fry, 2004).

Given the educational challenges faced by ELs, research has 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). The PLATO protocol for English language arts teaching, for example,

measures accommodations for EL students (Grossman, Loeb, Cohen, & Wyckoff, forthcoming). Other research has explored whether ELs benefit differentially in terms of math learning from having teachers with particular characteristics such as prior experience teaching English learners (Master, Loeb, Whitney, & Wyckoff, 2012). This research builds on prior studies that have found that on average teachers with more experience (Clotfelter, Ladd, & Vigdor, 2007; Harris & Sass, 2011; Kane, Rockoff, & Staiger, 2008; Nye, Konstantopoulos, & Hedges, 2004; Rice, 2003; Wayne & Youngs, 2003), content knowledge (Hill, Rowan, & Ball, 2005; Rockoff, Jacob, Kane, & Staiger, 2011), and particular types of preparation (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Ronfeldt, forthcoming) 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 FEPs 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 one student effective for all?

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.

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 reduced-price lunch eligibility, special education status, and whether the students are limited English proficient. In this paper, we define ELs as students in our dataset who are classified as ELs or have been within the prior three years. 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, 2, and 3. To answer research question 4, we append two teacher characteristics to the dataset: Spanish fluency and whether a teacher has a bilingual certification. 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 showing how this proportion varies by grade for the grades for which we have test scores. Between the 2003-04 and the 2009-10 school years the proportion of ELs remained fairly constant, ranging from 0.12 to 0.14, with a slight uptick in 2010-11 to 0.16. 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 FEPs.

Table 2 describes our sample at the student-, class-, and school-level, overall, for EL students and for FEP students. Not surprisingly, ELs are more likely than FEPs to be Hispanic and less likely to be black or white. Further, a higher percentage of ELs (81 percent) are eligible to receive free or reduced-price lunch compared to FEPs (65 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, roughly 60 percent of students in the average student's class were Hispanic and 13 percent were EL (this percent would be much higher were we only to look at 2010-11, the year in which the concentration of ELs was highest). As for teachers, on average 41 percent were fluent in Spanish and 5 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 FEP students, on average.

Methods

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

1) How much do teachers explain differences in learning among ELs and how much does this effectiveness differ from teacher effectiveness with FEPs?
2) To what extent are teachers who are effective with ELs also effective with FEPs and vice-versa?
3) How much of the variance in teacher effectiveness with ELs and FEPs as measured by value added occurs between schools versus among teachers in the same school?
4) 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.

This 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.

Using value-added modeling has several analytical strengths, not least of which is mitigating potential biases. Much of the confounding influence of unobserved student characteristics, which in turn influence estimates of teacher and school performance, can be reduced comparing students who start off with similar prior test scores. This approach

essentially controls for the student academic, family, and other characteristics that affected his or her prior performance. Additional controls for measured background characteristics and peer characteristics further mitigate the potential biases.

We calculate value-added estimates in the form of coefficients on teacher fixed effects used to predict student test score gains. For all of our teacher fixed effects models, we calculate value-added for ELs and FEPs separately in order to compare the estimates. Further, we only run models for teachers who have ten or more students in either category to ensure the estimates are based on a sufficient number of observations.

Specifically, we estimate a teacher fixed-effects model, as described by Equation 1, that predicts the test score gain between year *t-1* and year *t* for student *i* in grade *g* with teacher *j* in school *s* as a function of student (X_{ijst}), classroom (C_{jt}), and school (S_{st}) characteristics. Student-level covariates include time-varying characteristics such as whether the student qualifies for free or reduced price lunch, as well as stable characteristics such as race and gender. In addition to teacher fixed effects (δ_j), we also include year (γ_t) and grade (α_g) fixed effects. Finally, to estimate value added from one year to the next, we control for a vector of prior year test scores ($\mathbf{A}_{igjs(t-1)}$) in both math and reading. For simplicity, subscripts for academic subject are omitted, but we estimate the model separately for math and reading.

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

As is most often the case, our grade and year fixed-effects serve primarily as controls. The parameter γ_t captures any unobservable differences in test score gains that might be due to variance from year to year (for example, if a district-wide policy changed between years, that would be accounted for by γ_t). Similarly, the parameter α_g captures any inherent differences in test-score gains that occur from one grade to the next, such as a more difficult assessment being

used. Our student, classroom, and school characteristics, meanwhile, 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.

While Equation 1 is includes rich controls, the model could still suffer from bias. For instance, there could be student-level unobservable characteristics that influence the rate at which a student makes gains on an achievement test, such as innate ability, motivation, and family attitudes towards education. To help address these potential shortcomings, we estimate a model similar to Equation 1 that includes student fixed effects. In the specification with student fixed effects, we use achievement gains as the outcomes (rather than achievement in the current year as the outcome with lagged achievement as a predictor), and we no longer include stable student characteristics as controls because they are absorbed by the student fixed effect. In essence, including a student fixed effect controls for all unobservable time-invariant student characteristics that might influence learning and results in the comparison of a student's gains in one year to his or her mean gain across years.

The model including a student fixed effect theoretically does a better job of isolating the teacher effect because it removes the bias that could be due to nonrandom sorting of students based on time-invariant student characteristics. However, this specification does not necessarily remove bias from nonrandom sorting of students based on time-varying characteristics (Rothstein, 2009). Furthermore, the average standard errors of the teacher effects are higher than those obtained from the model using student covariates (McCaffrey, Sass, Lockwood, & Mihaly, 2009). The larger standard errors result from using many more degrees of freedom, which in turn explain much of the variation in teacher effects. Goldhaber and Hansen (2010) corroborate McCaffrey's assertion by pointing out that we commonly cannot reject the joint hypothesis that

the student fixed effects are zero, so by dropping them from the model, we can achieve more efficient estimation. Nonetheless, we perform our analyses using both specifications (results from the student fixed effects model are available upon request). Finally, 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 FEPs. We then use regression models to investigate which teacher characteristics are associated with higher gains. We provide a general description of the models and analytical approaches used below, while a complete list of covariates appears in Appendix 2.

Research Question 1: How much do teachers explain differences in learning among ELs and how much does this differ from teacher effectiveness with FEPs?

To answer this question, we compare the variances of the teacher fixed-effects estimates in math and reading for teachers of ELs and FEPs. We make this comparison for three different versions of each value-added estimate: the original, the "true" estimate, which backs out measurement error, and the shrunk estimate. 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. In essence, this approach helps account for the proportion of a teacher's value added that is due to measurement error. We produce the shrunk variance simply by taking the variance of the Bayesian shrunk estimates of the fixed effects described above.

We also use two statistical tests to determine whether the variances for ELs and FEPs are, in fact, significantly different. For the original and shrunk value-added estimates, we use the

Levene test for equality of variances. However, we cannot use the 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. Therefore, as a baseline to which we can compare the true variances, we also compare our estimates to true estimates for groups of randomly generated ELs and FEPs. 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 30 times so that we have a distribution of variances for random ELs and FEPs. We can then see where the variances for ELs and FEPs falls on the sampling distribution and we can see where the difference in variance between the two groups falls on the sampling distribution for the random differences.

Research Question 2: To what extent are teachers who are effective with ELs also effective with FEPs and vice-versa?

We address this question in two primary ways. First, we correlate value-added gains for teachers of ELs and FEPs separately for math and reading, as well as cross-tabulate EL and FEP value-added estimates by quintile. The former allows us to see how highly teacher value-added for ELs and FEPs correlates, and the latter gives an indication of how frequently teachers who are effective (or ineffective) with ELs are effective (or ineffective) with FEPs.

Second, we compare correlations between value-added estimates for ELs and FEPs to the same correlations from our randomly generated sets of ELs and FEPs. If the correlations are similar then the differences between a teacher's estimated value added with ELs and his/her estimated value added with FEPS can be attributed largely to measurement error. However, if the correlation between teachers value-added with ELs and FEPs is lower than between

randomly generated student groups it is likely that some teachers are differentially effective with ELs. We supplement with analysis with random effects estimates which, though more parameterized, allows for the direct measure of the correlation between the two groups.

Research Question 3: How much of the variance in teacher effectiveness as measured by value added occurs between schools rather than among teachers within the same school?

This section of the paper relies primarily on ANOVA estimates to compare within- and across-school variance in teachers' value-added. Specifically, we run ANOVA models by academic subject and school level using value-added estimates as the dependent variable and school indicator variables as the independent variable. We first look at the ANOVA results to see whether schools are a significant predictor of value-added for each group. In the event we can reject the null hypothesis of the F-test that there is no variance across schools, we use the adjusted R-squared values to determine how much of the variance in the estimated teacher value-added scores is accounted for by across-school differences for teacher effectiveness with ELs and for teacher effectiveness with FEPs.

Research Question 4: Can measured teacher characteristics help explain these differences in value-added?

To better understand what helps explain differences in value-added estimates, we regress teacher characteristics on year-to-year test score gains. Specifically, our models include covariates for a teacher's Spanish fluency and whether he or she has a bilingual certification. Our student achievement growth 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. In the base model, we include grade and year fixed

effects as controls, but otherwise include no fixed effects. The specification is detailed below, which includes the characteristic of interest, κ, and its interaction with EL status.

$$A_{igjst} = \mathbf{A}_{igjs(t-1)}\,\beta_1 + \mathbf{A}_{igjs(t-2}\,\beta_{2)} + \mathbf{X}_{it}\,\beta_3 + \mathbf{C}_{jt}\,\beta_4 + \mathbf{S}_{st}\,\beta_5 + \beta_6\kappa_j + \beta_7\kappa_j{*}EL + \gamma_t + \alpha_g + \varepsilon_{igjst} \quad (2)$$

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 another 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 FEP students within a teacher's classroom to investigate whether a teacher with specific attributes is more effective with either group. Note that in the teacher fixed effect model, the teacher characteristic of interest is omitted, as it is absorbed by the teacher fixed effect.

Findings

In this section of the paper, we discuss the findings from our analysis.

Research Question 1: How much do teachers explain differences in learning among ELs and how much does this effectiveness differ from teacher effectiveness with FEPs?

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 FEP. As discussed in the methods section, for each set, we calculate the standard deviation of: 1) the original value-added estimates, 2) the "true" value-added estimates from which we have subtracted measurement error, and 3) the Bayesian shrunk estimates. Our findings dovetail with those produced in other value-added research (Hanushek & Rivkin, 2010). Specifically, Hanushek and Rivkin (2010)

find a shrunk standard deviation in math value-added of .10 (we find .103). Overall, we also see that teacher quality varies the most at the elementary level with a shrunk estimated standard deviation of .128 for ELs and .153 for FEPs.

The "true" estimates are our best guess for the actual variance of value-added. While, in most cases, the estimates of variance are greater for teachers of FEPs than ELs, the differences are relatively small and do not appear to be different from what we would expect from a random draw from similar populations with equal variances. Our results say, for example, that 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 FEPs. To see 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 30 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 FEPs 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 FEPs separately, the difference in the variance estimates is similar to that of the fixed effects model for math, and slightly larger for reading.

The variances differ for the original and shrunk estimates, though neither of these are as accurate measures of variance than the true. Specifically, the variance is greater for ELs when using original estimates, but the reverse is true for shrunk estimates. This reversal is unsurprising given one would expect to find greater measurement error for ELs, an error the shrunk estimates attempt to address. To see whether these differences between ELs and FEPs

are significant (both for the original and shrunk estimates), we use Levene's test of equal variances. In the table, we use an asterice (*) to show that the standard deviations for the original and shrunk estimates are significantly different for ELs and FEPs as demonstrated by the Levene test.

Despite the significant differences we observe, the fact that the variances in teacher quality are greatest for ELs with the original estimates, greatest for FEPs with the shrunk estimates, and similar for the true estimates, our results generally suggest that the variances are quite similar for ELs and FEPs, 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 both types of students or if we use all teachers with available data, and find no observable differences in the results.

Research Question 2: To what extent are teachers who are effective with ELs also effective with FEPs and vice-versa?

In all of our models, we find high correlations between value-added for ELs and FEPs, though not as high as for randomly generated groups of students. In essence, teachers who are good with ELs tend to be good with FEPs and vice-versa, though some teachers are somewhat differentially 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 FEPs by quintile. First looking at the matrix for math, 59 percent of the teachers in the top quintile of value-added for FEPs are also in the top quintile of value-added for ELs. Of those teachers in the bottom quintile for FEPs, 50 percent are in the bottom quintile for ELs. These results suggest there is significant overlap in teachers who are effective with ELs and FEPs. Similarly, less than four

percent of teachers are either in both the top quintile for FEPs and in the bottom quintile for ELs or in both the bottom quintile for FEPs and in the top quintile for ELs. That is, 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 overall. Forty two percent of teachers in the top quintile for FEPs are in the top quintile for ELs and 35 percent of teachers in the bottom quintile for FEPs are in the bottom quintile for ELs. Only seven percent of teachers who are in the top quintile for FEPs are also in the bottom quintile for ELs and, again, only seven percent of teachers who are are in the top quintile for FEPs are also in the bottom quintile for ELs

Table 6 presents the correlations between value-added for ELs and FEPs by school level. In keeping with Table 4, the correlation for math is higher than for reading with the exception of high school. Also, when we randomly generate a group of ELs in the same proportion as is actually in a teacher's classroom, the correlation between value-added for ELs and FEPs is generally higher. 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 on of FEPs. In fact, across all school levels in math, we see that we would have been unlikely to draw to correlations as low as we did. For example, while we find a correlation of 0.61 between EL and FEP value-added, the average correlation from random draws is 0.68 with a standard deviation of 0.01. Thus, the difference between the actual and the random is approximately six standard deviations. In reading, the difference is not as clear with only a 1.63 standard deviation of the sampling distribution difference. These findings provide evidence that our imperfect correlations for ELs and FEPs are not due entirely

to measurement error. If correlations between real ELs and FEPs 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 FEPs compared to randomly generated ELs and FEPs suggests that there are likelier to be actual differences in value-added by group, though the differences are not great.[1]

Research Question 3: How much of the variance in teacher effectiveness with ELs and FEPs as measured by value added occurs between schools rather than among teachers within the same school?

Across all school levels, schools explain more of the differences between teachers for ELs than for FEPs. Table 7 shows the results of running an analysis of variance where value-added scores are predicted solely by the school in which a teacher teaches. Such a model's R-squared statistic indicates what percentage of the variance in value-added scores is explained by schools. For example, across all school levels, roughly 16 percent of the variance in value-added for ELs in math is between schools, whereas only 14 percent of the variance is between schools for FEPs (a statistically significant difference). The differences are greater for reading where 16 percent of the variance is between schools for ELs but only ten percent for FEPs. Generally, there is less variance explained by schools at the elementary level than at the middle or high school level. At the elementary level, seven percent of the variance is explained by schools for both ELs and FEPs in math. For ELs, seven percent is explained by schools for reading as well, but less is explained by schools for FEPs. In reading, in general, the variance of teacher value-

[1] 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.

added between schools is consistently higher for ELs than for FEPs. This result points to the importance of understanding schools' effects on ELs.

Research Question 4: What measured teacher characteristics help explain these differences in value-added?

Tables 8 and 9 show the results of student-level regression analyses that predict student achievement (in math and reading, respectively) as a function of teacher characteristics as described by Equation 2. The table gives results from three different models: 1) a model with no fixed effects, 2) a model that includes school fixed effects, and 3) a model with teacher effects. The coefficients presented are the regression coefficients from the interaction of EL with the relevant characteristic. Because student test scores are the outcomes, such a coefficient tells us what the achievement gap is between ELs and FEPs when they have a teacher with a particular characteristic. For example, at the elementary level, ELs experience a .07 standard deviation increase in math achievement over their FEP counterparts when they have a teacher who is fluent in Spanish, and a .11 standard deviation gain with a bilingually certified teacher.

In both math and reading, we see that all 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 more effective with ELs relative to FEPs. 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. Furthermore, bilingual certification is significant in all models for all school levels.

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 FEPs. That is, the variation in teacher effectiveness is generally as great for ELs as it is for FEPs. We also find that teachers who are effective with one group also tend to be effective with the other group. A good teacher does indeed tend to be a good teacher for all. 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 more a question of finding an effective teacher, not finding a teacher who specializes in English learners. The differential effectiveness of teachers with English learners is only a 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.

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 lightly touched on characteristics of teachers and schools 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. We clearly did not do as clean a job at identifying characteristics. Such work is beyond the scope of this paper because it requires both an understanding and discussion of teaching and learning and a convincing strategy for estimating causal effects. 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., & 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.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (Working Paper No. 17699). National Bureau of Economic Research.

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.

Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching.* ASCD.

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.

Ferguson, Ronald F. (2010, October 14). Student Perceptions of Teaching Effectiveness. Harvard University.

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

Fry, R. A. (2004). *Latino youth finishing college: The role of selective pathways* (pp. 1-40). Washington, DC: Pew Hispanic Center.

Goldhaber, D. & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions* (Working Paper #31). National Center for Analysis of Longitudinal Data in Education Research at the Urban Institute.

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

Grossman, P., Loeb, S. Cohen, J., & Wyckoff, J. (forthcoming). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added. *American Journal of Education*.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3), 466–479.

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.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, 26(1), 101–136.

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.

Lockwood, J. R., Daniel F. McCaffrey, Laura S. Hamilton, Brian M. Stecher, Vi-Nhuan Le, and Jose Felipe Martinez, (2007). "The sensitivity of value-added teacher effect estimates to different mathematics achievement measures," *Journal of Educational Measurement*, 44 (1), 47-67.

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*.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability. Monograph.* RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138.

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

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

Papay, J.P. (2011). Different Tests, Different Answers: The Stability of Teacher Value- Added Estimates across Outcome Measures. *American Educational Research Journal*, 48(1): 163-193.

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.

Perreira, K., Harris, K., & Lee, D. (2006). Making it in America: High school completion by immigrant and native youth. *Demography*, 43(3), 511–536.

Pianta, R. C., Paro, K. M. L., & Hamre, B. K. (2007). *Classroom Assessment Scoring System (Class) Manual, K-3*. Paul H. Brookes Publishing Company.

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

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

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. (forthcoming). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis*.

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.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8(3), 299-311.

Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects of teachers on future student academic achievement.

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	0.136	0.129	0.121	0.117	0.118	0.123	0.139	0.160
Percentage of ELs by grade								
Grade 3	0.182	0.187	0.155	0.168	0.165	0.184	0.199	0.255
Grade 4	0.131	0.107	0.117	0.106	0.127	0.125	0.160	0.183
Grade 5	0.125	0.127	0.119	0.112	0.102	0.110	0.127	0.153
Grade 6	0.134	0.135	0.135	0.102	0.094	0.098	0.104	0.101
Grade 7	0.127	0.118	0.104	0.109	0.098	0.091	0.119	0.115
Grade 8	0.125	0.127	0.104	0.098	0.099	0.107	0.103	0.110
Grade 9	0.110	0.091	0.113	0.098	0.108	0.099	0.109	0.128
Grade 10	0.111	0.100	0.095	0.094	0.091	0.096	0.102	0.110
Standardized math test scores	-0.662	-0.734	-0.750	-0.763	-0.731	-0.763	-0.755	-0.655
Standardized reading scores	-0.923	-0.995	-1.026	-0.996	-1.044	-1.032	-0.972	-0.931

Table 2

Race/ethnicity, Free or reduced price lunch, and Special Education, by EL status

	Overall		EL		FEP	
Student	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
White	0.09	-	0.02	-	0.10	-
Black	0.28	-	0.14	-	0.31	-
Hispanic	0.61	-	0.83	-	0.57	-
Free or reduced price lunch	0.67	-	0.81	-	0.65	-
Special education	0.19	-	0.13	-	0.20	-
Ever EL	0.13	-	1.00	-	0.00	-
Math Score	-0.17	0.98	-0.72	1.06	-0.08	0.94
Reading Score	-0.16	0.98	-0.99	1.03	-0.04	0.92
Class	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
White	0.09	0.28	0.02	0.15	0.10	0.29
Black	0.28	0.45	0.14	0.34	0.31	0.46
Hispanic	0.61	0.49	0.83	0.38	0.57	0.49
Free or reduced price lunch	0.67	0.47	0.81	0.39	0.65	0.48
Special education	0.19	0.40	0.13	0.34	0.20	0.40
Ever EL	0.13	0.34	1.00	0.00	0.00	0.00
Math Score	-0.19	1.11	-1.01	1.11	-0.10	1.07
Reading Score	-0.19	1.09	-1.34	1.00	-0.07	1.03
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	0.12	0.02	0.10	0.10	0.12
Black	0.28	0.32	0.14	0.30	0.31	0.33
Hispanic	0.61	0.31	0.83	0.31	0.57	0.32
Free or reduced price lunch	0.67	0.24	0.81	0.25	0.65	0.24
Special education	0.19	0.12	0.13	0.18	0.20	0.13
Ever EL	0.13	0.10	1.00	0.00	0.00	0.00
Math Score	-0.17	0.45	-0.72	0.66	-0.09	0.46
Reading Score	-0.17	0.42	-0.99	0.60	-0.04	0.44

Table 3

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

	EL			FEP			Difference	Random Difference	SDs Apart
Model 1	**Original**	**TRUE**	**Shrunk**	**Original**	**TRUE**	**Shrunk**	**TRUE**	**TRUE**	**TRUE**
All Grades									
Math	0.206*	0.153	0.103*	0.191*	0.171	0.125*	0.018	0.035 (0.010)	-1.6435
Reading	0.177*	0.110	0.063*	0.157*	0.127	0.079*	0.017	0.022 (0.006)	-0.7057
Elementary									
Math	0.238*	0.230	0.128*	0.226*	0.212	0.153*	0.018	-0.020 (0.021)	1.8254
Reading	0.204*	0.144	0.074*	0.177*	0.142	0.084*	0.002	0.000 (0.012)	0.1532
Middle									
Math	0.188	0.141	0.092*	0.181	0.163	0.109*	0.022	0.035 (0.011)	-1.1405
Reading	0.152*	0.083	0.045*	0.130*	0.094	0.053*	0.011	0.019 (0.011)	-0.7803
High									
Math	0.198*	0.149	0.094*	0.152*	0.128	0.073*	-0.012	0.028 (0.010)	-4.6987*
Reading	0.168*	0.104	0.061	0.149*	0.115	0.067	0.055	-0.002 (0.009)	1.3669

*Note. Numbers in table are reported in standard deviations. A * means a Levene test comparing standard deviations from the distributions of value-added estimates for ELs and FEPs found a statistically significant difference (.05) between the standard deviations of the groups.*

Table 4

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

EL VA	FEP 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 FEPs (reading)

EL VA	FEP 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 FEPs for real and randomly generated ELs						
	EL		**Randomly generated EL mean (SD)**		**Standardized Difference**	
	Math	**Reading**	**Math**	**Reading**	**Math**	**Reading**
All	0.61	0.44	0.68 (.01)	0.48 (.02)	6.15	1.63
Elementary	0.67	0.45	0.73 (.02)	0.48 (.03)	3.88	1.00
Middle	0.65	0.39	0.67 (.02)	0.45 (.03)	0.85	2.49
High	0.42	0.44	0.48 (.03)	0.47 (.03)	1.69	0.93

Note. Randomly generated EL mean and standard deviations based on 30 runs.

Table 7

Comparing across-school variance (ANOVA) for teachers of ELs and FEPs

	Math	Reading
	Adj. R-sq	**Adj. R-sq**
All Levels		
EL	0.160***	0.155***
FEP	0.144***	0.097***
Elementary		
EL	0.068*	0.072**
FEP	0.069***	0.030**
Middle		
EL	0.105***	0.083***
FEP	0.139***	0.079***
High		
EL	0.186***	0.102***
FEP	0.232***	0.022

*Note. *p<.1, **p<.05, ***p<.01.*

Table 8

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

Math	EL versus FEP Achievement Gap		
	No fixed effects	School fixed effects	Teacher fixed effects
All Levels			
Spanish Fluency * EL	0.04*	0.028*	0.034*
	(0.016)	(0.012)	(0.014)
Bilingual Certification * EL	0.012	0.017	0.043
	(0.037)	(0.036)	(0.044)
Elementary			
Spanish Fluency * EL	0.074*	0.056*	0.042~
	(0.033)	(0.026)	(0.025)
Bilingual Certification * EL	0.11~	0.134**	0.17**
	(0.062)	(0.048)	(0.063)
Middle			
Spanish Fluency * EL	0.04	0.029	0.048*
	(0.025)	(0.018)	(0.022)
Bilingual Certification * EL	-0.015	-0.032	-0.01
	(0.045)	(0.043)	(0.043)
High			
Spanish Fluency * EL	0.008	0.009	0.012
	(0.025)	(0.018)	(0.023)
Bilingual Certification * EL	0.154	0.132	0.767
	(0.535)	(0.532)	(0.725)

*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 9

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

Reading	**EL versus FEP Achievement Gap**		
	No fixed effects	**School fixed effects**	**Teacher fixed effects**
All Levels			
Spanish Fluency * EL	0.03*	0.022~	0.009
	(0.014)	(0.013)	(0.013)
Bilingual Certification * EL	0.082***	0.08***	0.063*
	(0.024)	(0.024)	(0.027)
Elementary			
Spanish Fluency * EL	0.043	0.013	0.019
	(0.028)	(0.026)	(0.027)
Bilingual Certification * EL	0.128**	0.128**	0.162***
	(0.05)	(0.046)	(0.039)
Middle			
Spanish Fluency * EL	0.026	0.026	0.008
	(0.021)	(0.017)	(0.021)
Bilingual Certification * EL	0.072*	0.08*	0.081*
	(0.033)	(0.033)	(0.035)
High			
Spanish Fluency * EL	0.019	0.012	0.000
	(0.022)	(0.022)	(0.02)
Bilingual Certification * EL	0.07~	0.058	-0.035
	(0.036)	(0.039)	(0.057)

*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[2]:

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

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 et al. 2006; Jacob & Lefgren 2005).

[2] 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
- Twice 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