DO HIGH-SCHOOL TEACHERS REALLY MATTER?

C. Kirabo Jackson

Working Paper 17722
http://www.nber.org/papers/w17722

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2012

I would like to thank David Figlio and Jon Guryan for helpful comments. All errors are my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Do High-School Teachers Really Matter?
C. Kirabo Jackson
NBER Working Paper No. 17722
January 2012
JEL No. H0,I20,J00

ABSTRACT

Unlike in elementary schools, high school teacher effects may be confounded with unobserved track-level treatments (such as the AVID program) that are correlated with individual teachers. I present a strategy that exploits detailed course-taking information to credibly estimate the effects of 9th grade Algebra and English teachers on test scores. I document substantial bias due to track-specific treatments and I show that traditional tests for the existence of teacher effects are flawed. After accounting for bias, I find sizable algebra teacher effects and little evidence of English teacher effects. I find little evidence of teacher spillovers across subjects.

C. Kirabo Jackson
Northwestern University
School of Education and Social Policy
2040 Sheridan Road
Evanston, IL 60208
and NBER
kirabo-jackson@northwestern.edu

Do High School Teachers Really Matter?

By C. KIRABO JACKSON[1]
JAN 1, 2012
Northwestern University, IPR, and NBER

Unlike in elementary schools, high school teacher effects may be confounded with unobserved track-level treatments (such as the AVID program) that are correlated with individual teachers. I present a strategy that exploits detailed course-taking information to credibly estimate the effects of 9th grade Algebra and English teachers on test scores. I document substantial bias due to track-specific treatments and I show that traditional tests for the existence of teacher effects are flawed. After accounting for bias, I find sizable algebra teacher effects and little evidence of English teacher effects. I find little evidence of teacher spillovers across subjects. (JEL I21, J00).

There is consensus among policy-makers and researchers that teachers are the most important component of schools. This conclusion is based, in large part, on studies linking individual elementary school teachers to their students' test-scores. A variety of studies using different elementary school student populations show that a one standard deviation increase in teacher quality leads to between one-tenth and one-fifth of a standard deviation increase in math and reading scores (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004; Kane & Staiger, 2008), and recent findings indicate that effective elementary and middle-school teachers are associated with improved long run outcomes (Chetty, Friedman, & Rockoff, 2011).

There is much less consensus regarding high school teachers. Using similar methodologies in high school as used in elementary school, a few researchers find similar effects on math (Aaronson, Barrow, & Sander, 2007) and reading (Koedel, 2008) achievement. However, because students in elementary school are typically exposed to only one teacher and are all in the same academic track while those in secondary schools are exposed to several different teachers and are placed into different tracks, value-added methodologies designed for elementary school teachers may be inappropriate for measuring teacher quality in high school. In a high school setting, *even with random assignment of students to teachers,* if different teachers teach in different tracks, and students in different tracks are exposed to different treatments, there will be bias due to "track treatment effects." These "track treatment effects" may arise due to

[1] I would like to thank David Figlio and Jon Guryan for helpful comments. All errors are my own.

other teachers (e.g. students who take algebra I with Mr. Smith take physics with Mr. Black who has a direct effect on students' algebra scores), the content of other courses (e.g. students in the high-ability track take Algebra I with Mr. Smith and also take physics which has a direct effect on Algebra I scores), or explicit track level treatments (e.g. students in the high ability track take Algebra I with Mr. Smith and are also part of the AVID[2] program that teaches study-skills and time-management skills which have a direct effect on algebra scores). As such, in high schools there is *both* possible selection bias due to non-random placement of students to tracks (Koedel & Betts, 2009; Rothstein, 2010), *and* omitted variables bias due to track specific treatments. This second source of bias creates additional challenges to identifying teacher effects in high school and has not been accounted for in the existing literature. As such, existing studies may be misleading about the importance of high school teachers— raising important questions about the efficacy of policies based on teacher value-added in high school such as merit pay or retaining and rewarding effective teachers.

To address these concerns, I (a) propose and employ a strategy to estimate the effects of math and English teachers on test scores in 9th grade that exploits detailed course-taking information to remove both sources of bias, (b) analyze the extent to which historically estimated teacher effects predict teacher effectiveness in a high school setting, and (c) present a quasi-experimental test to show that the estimated teacher effects are not driven by student selection.

To address the unobserved track-level treatment problem, because I can observe all the courses students take *and* the level of instruction within particular courses, I can control for the unique set of courses (and their levels) taken by students so that all comparisons are made among students who are at the same school and in the same academic track. Comparing the outcomes of students with different teachers at the same school taking the same course and in the same track removes the influence of any school-by-track level treatments that could confound comparisons of teachers who teach in different tracks. Making comparisons among students in the same track and school also removes bias due to sorting or selection into schools, tracks, and courses.

In such models, variation comes from comparing the outcomes of students in the same track and school but are exposed to different teachers either due to (a) changes in the teachers for

[2] According to the Advancement Via Individual Determination (AVID) Mission Statement "AVID offers a rigorous program of instruction in academic “survival skills” and college level entry skills. The AVID program teaches the student how to study, read for content, take notes, and manage time. Students participate in collaborative study groups or tutorials led by tutors who use skillful questioning to bring students to a higher level of understanding."

a particular course and track over time, or (b) schools having multiple teachers for the same course and track in the same year. Because personnel changes within schools over time may be correlated with other improvements within schools, I estimate models that include school-by-year fixed effects. The remaining concern is that comparisons among students within the same track may be susceptible to selection bias. I argue that most plausible stories of student selection involve selection to tracks or courses rather than to teachers *within* tracks or courses, so that teacher effects based on within-school-track variation should not be confounded by selection. However, to assuage this concern, I show that conditional on track and course fixed-effects, teacher assignments are orthogonal to observable student characteristics. In addition, I exploit the statistical fact that any selection within tracks will be eliminated by aggregating the treatment at the track level, and show that over time cohorts in tracks exposed to higher average estimated pre-sample teacher value-added in the track have better outcomes on average.

While the primary focus of the paper is to estimate high school teacher effects, this paper makes a methodological contribution to the broader teacher quality literature. I show that the typical test for the existence of teacher effects is severely biased. It is common practice to use an F-test to determine whether there is excess heterogeneity in test scores across teachers beyond what would be expected due to sampling variability. It is well known that transitory classroom-level shocks that affect student test scores but are not due to teaching ability (e.g. a dog barking on exam day or a flu outbreak in a classroom) cause raw value-added estimates to overstate the variability of teacher quality (Kane & Staiger, 2008). It is less well known that classroom-level disturbances also cause the commonly used *F*-test to erroneously reject the null hypothesis of equality of effects even when it is true. With classroom-level shocks, the *F*-test for excess heterogeneity in test scores across teachers *is not a test for the equality of teacher effects*, but is a test of equality of the average classroom effects for each teacher. In the extreme case, where teachers are observed in only one classroom, one cannot distinguish a classroom shock from a teacher effect (e.g., one cannot distinguish a good teacher from a lucky teacher).[3] Because many studies use the *F*-test to show the existence of teacher effects, the extant literature may overstate the degree to which one can be confident that teachers affect student outcomes.

[3] In fact, researchers often use the term classroom effect and teacher effect interchangeably. While a teacher effect will certainly contribute to the classroom effect, the converse is not true.

To document the bias associated with the *F*-test, I simulate hypothetical data under the assumption of no teacher effects and show how the presence of classroom shocks affects the likelihood of erroneously rejecting the null of no teacher effects. I then present an unbiased test based on the covariance of mean classroom residuals for the same teacher over time.

While I find little evidence of bias due to student sorting, I find evidence of substantial bias due to omitted track-level treatments. Specifically, the estimated variability of teacher effects falls by over half with the inclusion of track fixed effects. This decline is not driven by teacher sorting. Based on covariance across classrooms for the same teacher, the 95% confidence interval for the variance of algebra teacher effects is [0.11σ , 0.24σ], while that for English teachers is [-0.09σ ,0.212σ]. As such, while one can rule out zero persistent algebra teacher effects, one cannot rule out zero teacher effects for English. Estimated value-added (from out of sample) is unrelated to observed student characteristics that predict test scores. Also, I present a quasi-experimental test based on aggregating the treatment at the track level showing that these estimated effects are not biased due to student sorting in unobserved dimensions.

Even though the biased *F*-test indicates large teacher effects for both subjects, the unbiased test proposed rejects no effect for algebra but fails to reject no effect for English. In addition, while the *F*-test would indicate large teacher spillover effects across subjects (as found in other studies of high school teachers[4]), the unbiased test fails to reject the null of no spillovers. This suggests that the evidence of spillovers across subjects found in other studies may have been due to the biased *F*-statistic test rather than being a real phenomenon.

To assess the predictive ability of these estimates, I set up a quasi-experimental design where I compare the outcomes of students between 2008 and 2010 who have teachers with high and low value-added from 2005 through 2007. The predictive ability of value-added estimates is much lower in the high school setting than in the elementary school setting. Specifically, while studies of elementary school teachers find that a teacher who raises tests scores by 1σ in pre-sample improves student outcomes by about 0.4σ in math and reading, I find that a high school teacher who raises tests scores by 1σ in pre-sample only improves algebra test scores by 0.1σ and has no effect on English test scores. This suggests that the properties of teacher value-added in high school are quite different from those in elementary school such that polices aimed at using teacher test score value-added in high school may not yield dramatically improved results.

[4] Aaronson, Barrow, and Sander (2007); Koedel (2008).

This paper makes a few contributions to the teacher quality literature. This is the first study to highlight the unique identification challenges in a high school setting and to show that track-level treatments are important. I also present a new methodology that policy-makers and researchers can use to credibly identify high school teacher effects. Also, this is the first study to show that persistent teacher test-score effects do not exist for all subjects. This paper also highlights limitations of the statistical tests used in the extant literature and presents an unbiased test for the existence of teacher effects that can be used in both the elementary and high school context. Finally, this paper demonstrates that there are greater challenges in using test-score based measures of quality for high school teachers in personnel decisions than for elementary school teachers (particularly for English and perhaps other subjects also).

The remainder of the paper is as follows: Section II describes the data used, Section III details the empirical framework, Section IV details and the identification strategy, Section V presents the main results, robustness checks, and specification checks, and Section VI concludes.

II Data:

This paper uses data on all public middle- and high school students in North Carolina from 2005 to 2010 from the North Carolina Education Research Data Center.[5] The student data include demographic data, transcript data for all courses taken, middle-school achievement data, end of course scores (high school) for Algebra I and English I, and codes allowing one to link students' end of course test-score data to individual teachers who administered the test.[6] Because the teacher identifier listed is not always the student's teacher, I link these data to detailed personnel records and teaching activities and remove all student and teacher records that are not associated with a regular classroom teacher who teaches Algebra I or English I. Because English I and Algebra I are the two tests that have been the most consistently administered over time, I limit the analysis to students who took either the Algebra I course or the English I course. Over 90 percent of all 9th graders take at least one of these courses so that the resulting sample is representative of 9th graders as a whole. To avoid endogeneity bias that would result from

[5] These data have been used by Clotfelter, Ladd, and Vigdor (2010) to look at the effect of high school teachers qualifications on student test scores. They uses variation across subjects within the same student to identify the effect of teacher characteristics on student outcomes. This methodology does not allow one to identify individual teacher effects and does not address the bias due to track specific treatments.

[6] There are also end of course scores for History and Science. However, because exams are not given in all years for these exams I limit the analysis to students who took English I or Math I.

teachers having an effect on repeating ninth grade, the master data is based on the first observation for when a student is in ninth grade. Summary statistics are presented in Table 1.

The data cover 377,662 ninth grade students in 619 secondary schools in classes with 6538 English I teachers, and 6215 Algebra I teachers. While roughly half of the students are male, about 60 percent are white, 29 percent are black, 6 percent are Hispanic, 2 percent are Asian, and the remaining 1 percent are Native American mixed-race or other. About 3.8 percent of students have the highest parental education level (i.e. the highest level of education of the student's two parents) below high school, 27.7 with a high school degree, 8.5 percent with a junior college or trade school degree, about 19 percent with a four year college degree or greater (41 percent of observations in the dataset have missing parental education and are coded as such). About 10.8 percent of students are of limited English proficiency. The achievement data have all been normalized and standardized to be mean zero with unit variance for each cohort and test. Mean incoming 7th and 8th grade test scores in the 9th grade sample are approximately one-tenth of a standard deviation higher than that of the average in 7th or 8th grade. This is because the sample of 9th grade students is less likely to have repeated a grade and to have dropped out of the schooling system.

The Measure of Track:

One of the most important variables in the analysis is the measure of academic track and the definition of a school-track. Even though schools may not have explicit labels for tracks, they may practice de-facto tracking by placing students into distinct groups of courses. While there are hundreds of courses that students can take (including special topics, and reading groups) there are 20 courses will make up over 70 percent of all courses taken in 9th grade and some combination of these 20 courses would make up more than 75 percent of all courses for most students. I list these courses in Table 2. Of these twenty courses, 10 are elective courses and 10 are core academic courses (shown in bold). English I is the most common academic course taken such that 89.2 percent of 9th graders take English I, followed by World History that 84 percent of students take, Earth Science that 63 percent of students take, and algebra I that 51 percent of students take. Other common academic courses, that fewer than 50 percent of students take, include Art, Pre-Algebra, Biology, Introduction to Algebra, Basic Earth Science, and Spanish I.

A key detail in these data is also the level of the course taken. Even among students taking Algebra I or English I courses there are three different levels of instruction (advanced, regular, and basic) so that not all students who take Algebra I or English I are in the same academic track. As such, I exploit the richness of the data and take as my measure of a school-track, the unique combination of the 10 largest academic courses, the level of algebra I taken, and the level of English I taken in a particular school. As such, all students who take the same number and set of courses, and the same level of English I and Algebra I courses *at the same school* are in the same school-track. Students who take the same courses at different schools are in different school-tracks, students at the same school who took either a different number of courses or at least one different course are in different school-tracks, and students at the same school who took the same courses but took Algebra I or English I at different levels are in different school-tracks. Defining tracks at the school-by-course-by-level level allows different schools to have different selection models and treatments for each track.

While one may worry that some tracks have too few students to allow for identification, because many students pursue the same course of study only 3.7 percent of all students are in singleton tracks, 60 percent are in school-tracks with more than 50 students, and the average student is in a school-track with 117 other students. There are 18,226 non-singleton school-tracks across 726 schools. I present the same summary statistics for the data aggregated at the school-by-year level and the school-track-by-year level in Table 1. Comparing the standard deviations of the variables provides some indication of sorting into tracks. The standard deviation of reading and math scores in 8th grade are about 0.95 for the student level data and track level data. Because track-by-school level data are more aggregated, if students were randomly assigned to schools and tracks the standard deviation of the school by track means would be smaller than that of the overall data. Based on size of tracks, the standard deviation of the track by school level means would be approximately 2.5 times smaller under random assignment. However, the standard deviation of the individual data is similar to that of the school-track level means, suggesting that students are systematically grouped into tracks by incoming achievement in a manner that increases incoming test score dispersion across tracks. One can see a similar pattern for other covariates; students are systematically grouped into tracks by ethnicity, parental education, and LEP status in a manner that increases differences across tracks. For the outcomes, Algebra I and English I scores, the standard deviation of the school-by-track-by-year level data are greater than the individual level data— indicating that something at the track level increases dispersion of these outcomes. This could be due to teacher quality, sorting into tracks, or track level treatments. Note that much of this evidence of tracking may occur at the school level rather than at the track level within schools.

III Empirical Framework under Tracking:

The main objective of this paper is to estimate the effect of individual math and English teachers on 9th grade students' test scores. This entails comparing the outcomes of students who have one teacher to the outcomes of students who have another teacher. In a high school setting obtaining consistent estimates of teacher effects can be challenging for two reasons: First, students may select to tracks in unobserved dimensions that would lead to selection-bias; Second, high school students are often placed into tracks (groups of courses) such that taking a class with a particular teacher means taking a particular course, which may be associated with taking a particular set of other courses, being counseled in a different way and being exposed to different peers (in other classes). These "treatments" associated with tracks could have an independent effect on outcomes and may confound estimated individual teacher effects.

Specifically, for any school-track *c*, if students select to tracks within high schools then there may be unobserved determinants of outcomes that are correlated with a student's track. Also, with track-specific treatments ($P|c$) there are additional unobserved determinants of student outcomes associated with student's school-tracks. With additive separability of inputs in the production of student achievement where lagged achievement is a sufficient statistic for the history of family school and student inputs, one can write student achievement as [1] below.

[1] $$Y_{ijcy} = A_{iy-1}\delta + X_{iy}\beta + I_{ji}\cdot\theta_j + \pi(P|c) + \varepsilon_{ijcy}$$

Here, Y_{ijy} is the outcome of student *i* with teacher *j* in school-track *c* in year *y*, A_{iy-1} is incoming achievement level of student *i*, X_{iy} is a matrix of student level covariates obtained in 8th grade (including parental income, Limited English Proficiency status, ethnicity, and gender), I_{ji} is an indicator variable equal to 1 if student *i* is in class with teacher *j* and equal to 0 otherwise, θ_j is a teacher fixed effect, ($P|c$) is a treatment specific to students in school-track *c*, and ε_{ijcy} is the idiosyncratic error term. When the school-track is unobserved, the conditional expectation of teacher effect $\hat{\theta}_j$ is given by [2] below.

[2] $$\mathrm{E}(\hat{\theta}_j \mid I_j, \mathrm{X}, \mathrm{A}) = \theta_j + (\sigma_{\mathrm{I}_{ji},c} / \sigma^2_{\mathrm{I}_{ji}})[\pi(P \mid c) + E(\varepsilon_{ijcy} \mid \mathrm{c})].$$

As long as teachers are not randomly assigned to tracks (i.e. $\sigma_{\mathrm{I}_{ji},c} / \sigma^{2}_{\mathrm{I}_{ji}} = 0$) value-added estimates are only consistent if there is no selection to tracks (i.e. E[ε_{ijcy} |c]= E[ε_{ijcy}]) and there are no other treatments associated with being in track (i.e. P=0). While previous analyses have accounted for the first source of bias, none have accounted for the second. Also, failing to account for these sources of bias likely lead one to overstate the variability of teacher quality because the estimated teacher effects will also include the effect of the track-level treatments.

Without detailed information on exactly what treatments *P* are associated with each school-track *c*, and exactly what characteristics of students ε_{ijcy} are associated with each school-track *c*, it is difficult to control for these sources of bias directly. However, one can remove *both* the influence of any track specific treatments and the effects of selection across tracks by making inferences within groups of students *in the same track at the same school*. In a regression context, if one can observe track placement, this is achieved by including I_{ci}, an indicator variable equal to 1 if student *i* is in school-track *c* and 0 otherwise. This leads to [3] below.

[3] $$Y_{ijcy} = A_{iy-1}\delta + X_{iy}\beta + I_{ji}\cdot\theta_j + I_{ci}\,\theta_{\mathrm{c}} + \varepsilon_{icjy}$$

By conditioning on school-tracks, one can obtain consistent estimates of the teacher effects θ_j as long as there is no selection to teachers *within* a school-track.[7]

Sources of identifying variation:

Because the main models include school-by-track fixed effects, teacher effects are identified by comparing the outcomes of students at the same school in the same track but who have different teachers. In these models, identification of teacher effects comes from two sources of variation; (1) comparisons of teachers at the same school teaching students in the same track *at different points in time*, and (2) comparisons of teachers at the same school teaching students in the same track *at the same time*. To illustrate these sources of variation, consider the simple case illustrated in Table 3. There are five tracks A, B, C, D and E in a single school. Each track is defined by the school, the academic courses, and the level of the Algebra I and English I class taken. There are four math teachers at the school at all times but the identities of the teachers change from year to year due to staffing changes.

The first source of variation is due to changes in the identities of Algebra I and English I teachers over time due to staffing changes within schools over time. For example, between 2000

[7] Note: In expectation, the coefficient on the school-track indicator variable is $\pi(P|c)$+E[ε_{iy} |c]. This reflects a combination of *both* the unobserved treatment specific and selection to school-track c.

and 2005 teachers 3 and 4 were replaced with teachers 5 and 6. Because, teachers 3 and 6 both teach in tracks C and D (in different years) one can estimate the value-added of teacher 3 relative to teacher 6 by comparing the outcomes of students in tracks C and D with teacher 3 in 2000 with those of students in tracks C and D with teacher 6 in 2005. Similarly, because teachers 4 and 5 both teach in tracks B and E (in different years) one can estimate the value-added of teacher 4 relative to that of teacher 5. To account for any mean differences in outcomes between 2000 and 2005 that might confound comparisons within tracks over time (such as school-wide changes that may be coincident with the hiring of new teachers), one can use the change in outcomes between 2000 and 2005 for teachers 1 and 2 (who are in the school in both years) as a basis for comparison. In a regression setting this is accomplished with the inclusion of school-by-year fixed effects (Jackson & Bruegmann, 2009). This source of variation is valid as long as students do not select across cohorts (e.g. stay back a grade or skip a grade) or schools in response to changes in Algebra I and English I teachers. In Section IV.3, I show that differences in teacher value-added across cohorts are unrelated to observable student characteristics.

The second source of variation comes from having multiple teachers teaching the same course in the same track at the same time. In the example above, because both teachers 2 and 4 taught students in track B in 2000 one can estimate the value-added of teacher 2 relative to that of teacher 4 by comparing the outcomes of teachers 2 and 4 among those students in track B in 2000. Similarly, because both teachers 2 and 5 taught students in track B in 2005 one can estimate the value-added of teacher 2 relative to teacher 5 by comparing the outcomes of teachers 2 and 5 among those students in track B in 2005.[8] Because this variation comes from comparing teachers from the same track at the same school at the same time, the key identifying assumption is that while students may select to tracks, students do not select to individual teachers within tracks. In section IV.3 I show that differences in teacher value-added within school-tracks are unrelated to differences in observable student characteristics, and in section V.2, I show that the main findings are not driven by student selection within tracks.

[8] **NOTE:** One can identify all teacher effects for algebra I so long as there is sufficient overlap of courses and tracks. In this example, teacher 2 can be compared to 4 within track B in 2000, and compared to 3 within track C in 2000. Similarly, 4 can be compared to 5 within tracks B and E across 2000 and 2005, and 3 can be compared to 6 within tracks C and D across 2000 and 2005. As such, 2 can be directly compared to 3 and 4, and indirectly to 5 and 6 so that all teachers who teach Algebra I can be compared to each other.

IV Identification Strategy:

IV.1 Testing for Teacher Effects Using the F-test

Most studies on the effect of teachers on student outcomes estimate value-added models akin to equation [3] and report the *p*-value associated with the F-statistic for the test of the null hypothesis that all of the teacher effects, $\theta_j s$, are equal. Typically, researchers conclude that there are real teacher effects if the computed *F*-statistic is greater than the *F*-ratio associated with some critical value α. Formally, where there are N student observations, *K* teachers, and n_j students with teacher *j*, it is common practice to conclude that θ_js are not equal if [4] is true.

[4]
$$\text{F}=\frac{\sum_{i} n_j(\bar{Y}_j-\bar{Y})^2/(K-1)}{\sum_{ij}(Y_{ij}-\bar{Y}_i)^2/(N-K)}=\frac{\text{accross-teacher variance}}{\text{within-teacher variance}}>F_{\alpha}(K-1,N-K).$$

The *F*-statistic is the ratio between the across and within teacher variance. If the variance of the teacher fixed effects is large relative to the variance of the within-teacher variability in outcomes, then it would imply that there is something systematic occurring at the teacher level that explains variability in outcomes (test scores). I argue that in the presence of unaccounted for classroom-level shocks, the *F*-test commonly used is a severely biased test for the null hypothesis of no teacher effects and will tend to over-reject even when the null is true. These classroom-level shocks are those idiosyncratic shocks that affect the outcomes of all students in the classroom but are not accounted for with teacher fixed effects, observed student characteristics or even student fixed effects. Random shocks that affect test scores but are not due to persistent teacher quality could arise due to numerous unobserved factors such as the teacher falling ill during the school year, a flu outbreak in the classroom, a dog barking outside the classroom on test day, poor lighting, or a noisy boiler under the classroom. I demonstrate that adjusting for such shocks with clustering and aggregation do not solve the problem.

The intuition for this bias is straightforward. One can think about this problem of over-rejecting by considering the hypothesis that the *F*-test evaluates. The *F*-test tests for whether the mean of the classroom-level outcomes for each teacher is different from zero. When there are no classroom-level errors, this is equivalent to testing for whether the teacher effects are equal to zero. However, when there are classroom-level errors, this test no longer tests the hypothesis that the teacher effects are zero, but tests the null hypothesis that the sum of the teacher effects and the means of the classroom errors for each teacher are equal to zero. If classroom-level errors are normally distributed with variance σ_c, *even if all teacher effects are equal to zero*, the mean of the classroom-level residuals will be normally distributed with variance $\frac{\sigma_c}{n_{cj}} + \frac{Var(\varepsilon_{icjy})}{n_j}$, where n_{cj} is the number of classes observed with teacher *j*, and n_j is the number of students observed with teacher *j*. This equation makes explicit that when the variance of the classroom-level errors is large or the number of classrooms per teacher is small, the likelihood of observing a mean teacher level residual that deviates from zero even when there are no actual teacher effects is non-trivial. The problem this causes is best illustrated by the extreme case where each teacher is observed in only one classroom. In this scenario, if there is a large teacher level mean residual, one will correctly reject the null hypothesis of no classroom effects, but might interpret this as evidence of the existence of teacher effects. This example illustrates that one should not interpret the *F*-test as a test for the existence of teacher effects. I illustrate this bias below.

To get a sense of the behavior of the commonly used *F*-test, I simulated hypothetical data with no teacher effects but in which there are classroom-level disturbances. I created a dataset with 200 teachers each observed in 5 classrooms with one classroom per year. I ran 1000 replications of a regression of a mean zero unit variance normally distributed outcome Y on the set of teacher indicator variables where there are no teacher effects but there are normally distributed classroom-level effects with varying dispersion. Figure 1 shows the distribution of the *p*-values associated with the *F*-tests of equality of teacher indicator variables across the 1000 replications for classroom-level disturbances of different sizes. If the test is unbiased, the *p*-values should follow a uniform distribution centered around 0.5.

When there are no classroom-level errors (SD=0.00) the distribution of *p*-values appears to be roughly uniform. However, with just a small classroom-level error (sd=0.02) the likelihood of a *p*-value smaller than 0.1 is about 20 percent (twice as likely as should be if the test were unbiased) and with a modest classroom-level error (sd=0.05) the likelihood of a *p*-value smaller than 0.1 is 84 percent. Given that non-persistent components of teacher quality are about as large as teacher effects themselves, the variance of the classroom-level effects are likely closer to 0.15sd. At this level, even with no teacher effects the *F*-test is almost guaranteed to reject the null hypothesis of equality of the teacher effects.

One may think that clustering the standard errors at the teacher or class level would address this problem, but this is not the case. In fact, clustering exacerbates the problem such that even with no classroom-level errors one rejects the null hypothesis of no effect more than 80 percent of the time when the standard errors are clustered at either the classroom or teacher level. Figure 2 shows the mean *p*-value for different sizes of the classroom-level errors. If the tests were unbiased, the *p*-values should be close to 0.5. Even with a small classroom-level error the mean *p*-values are much smaller than 0.5 when one clusters errors to account for dependence within classes or teachers. This indicates that the *F*-test as commonly used is inappropriate for testing for teacher effects and motivates my proposal of an alternate statistical test.

IV.2 Empirical Strategy

While one approach to estimating the importance of teacher quality is to estimate equation [3] and compute the variance of the estimated teacher effects $\hat{\theta}$, this will overstate the variance of true persistent teacher quality because teacher effects are estimated with error due to sampling variation and there are classroom-level disturbances. As such, to estimate the importance of individual teachers on student outcomes I follow Kane and Staiger (2008) and Jackson (2010).[9] Specifically, I estimate equation [3] without teacher indicator variables, and compute the covariance of mean teacher-by-year level residuals for the same teacher over time as my estimate of the variance of the *persistent* component of teacher quality that is observed across years. Specifically, in the first stage I estimate equation [5] below.

$$[5] \qquad Y_{icjy} = A_{iy-1}\delta + X_{iy}\beta + X^*_c\pi_c + I_{ci}\,\theta_{\mathrm{c}} + \theta_{\mathrm{sy}} + \varepsilon^*_{ijcy}$$

The key conditioning variable is $I_{ci,}$ an indicator variable denoting the school-track *c* (defined at the school-by-course-group-by-course-level) of student *i*. A_{iy-1} is the third order polynomial of incoming math and English achievement of student *i*. To address concerns about dynamic tracking, I include math and reading test scores from both 7th and 8th grade (two lags of achievement). X_i is a matrix of additional student covariates such as parental education, ethnicity, gender, and LEP status. I also include the mean incoming test scores *and characteristics* of other students in the classroom X^*_c. To account for school-level time effects (such as the hiring of a new school principal) that would affect all students in the school, I also include school-by-year fixed effects θ_{sy}. Because this model does not include teacher indicator variables, the error term includes the teacher effect so that $\varepsilon^*_{ijcy} = \theta_j + \varepsilon_{ijcy}$.

[9] This procedure is also used in (Jackson, 2011) and (Jackson, 2009).

In the second stage, I compute mean residuals from [5] for each teacher in each year $n_{jy}^{-1}\sum_{i=1}^{n_{jy}} e^*_{ijcy} \equiv \theta_j + \overline{e}_{jcy}$, where n_{jy} is the number of students in class with teacher j in year y. To compute the variance of the persistent teacher quality, I compute the covariance of mean residuals for the same teacher in years y and year y-1. If the non-persistent error components for each teacher e_{jcy} are uncorrelated over time (recall that the model includes school-by-year fixed effects) and uncorrelated with teacher quality, the covariance of mean residuals for the same teacher over time is a consistent measure of the true variance of persistent teacher quality. If $Cov(\theta_j, \overline{e}_{jcy}) = Cov(\theta_j, \overline{e}_{jcy-1}) = Cov(\overline{e}_{jcy}, \overline{e}_{jcy-1}) = 0$ then $Cov(\varepsilon^*_{ijcy}, \varepsilon^*_{ijcy-1}) = \text{var}(\theta_j)$.

The test for the existence of persistent teacher quality in this framework effects is simply the p-value associated with the null hypothesis that the residuals for one year are uncorrelated with residuals from another for the same teacher. One can test this by running a regression of a teacher's mean residuals in year t on her mean residuals in year t-1 and then implementing a t-test for the coefficient on the mean lagged residuals. Even in the presence of large classroom-level disturbances, as long as classroom-level disturbances are uncorrelated over time this covariance based tests will be an unbiased test for the existence of persistent teacher quality effects. To illustrate the unbiasedness of this proposed test, I show how the covariance test performance on the same simulated data discussed in section IV.1 where the F-test was found to be problematic.

Figure 3 shows the distribution of the p-values associated with the covariance tests across the 1000 replications for classroom-level disturbances of different sizes. If the test is unbiased, the p-values should follow a uniform distribution centered around 0.5. As one can see, irrespective of the size of the classroom-level errors, the p-values follow a uniform distribution, so that this test is robust to idiosyncratic classroom-level disturbances and will be an unbiased test for the existence of teacher quality effects. As such, in addition to presenting the estimated covariance, I also present the results from tests for persistent teacher quality effects. It is worth noting that while aspects of teacher quality may be transitory, from a policy perspective, it is the persistent component that is important for predicting future teacher performance.

IV.3 Tests for Selection and Sorting Into Tracks and Within Tracks.

Before presenting the main results I test the extent of tracking following Aaronson, Barrow, and Sander (2007) and Koedel (2008) who assess the extent to which students may be sorted based on test scores or test score gains. I calculate mean within-teacher-year student test-score dispersion (i.e. the average across all teacher-years of the standard deviation of test scores computed for each teacher's classroom in a given year) for the observed teacher assignments and compare that to the mean within-teacher student test score dispersion for other counterfactual teacher assignments. Table 4 displays the average within-teacher-year standard deviation of 8th grade scores, 7th grade scores, and test-score growth between 7th to 8th grade for both math and reading. I present the actual within teacher-year test score dispersion, what one would observe with full student sorting (of the variable) within schools, full student sorting across all classroom and schools, random student assignment within schools and finally random student assignment across all classrooms and schools.

Comparing the actual test score dispersion to that when students are perfectly sorted across teachers within their school reveals that a within-school sorting mechanism would reduce the within-teacher-year standard deviation to between 25 and 30 percent of the actual observed within-teacher-year standard deviation, depending on the exact test score and subject. In contrast, the actual within-teacher-year test score dispersion is between 88 and 100 percent of what one would observe under random assignment of students to classrooms within schools, depending on the exact incoming achievement measure used. This is similar to findings in other studies, and suggests that tracking based on incoming achievement does exist, but may be relatively small.[10]

To provide further evidence of bias due to sorting, I determine if teacher value-added is correlated with observable student characteristics. Specifically, I estimate teacher value-added using data from 2005 through 2007[11], while computing predicted outcomes for students in 2008 through 2010. The predicted outcomes are fitted values from a linear regression of the outcomes on all observable student characteristics.[12] I then run a regression of standardized normalized teacher effects from 2005-2007 data on predicted outcomes from 2008-2010 (with school and year fixed effects) to see if students who are likely to have better outcomes (based on observed characteristics) tend to be assigned to teachers who had better or worse than average value-added

[10] In a Monte Carlo simulation of the mean test score dispersion under random assignment, none of the 200 replications yielded dispersion levels as low as that observed. As such, there is clearly some tracking in these data.

[11] I compute Empirical Baye's estimates following Kane and Stagier (2008). See appendix note 1.

[12] As discussed in Jackson (2010), this is a more efficient and straightforward test of the hypothesis that there may be meaningful selection on observables that estimating the effect of the treatment on each individual covariate. This is because (a) the predicted outcomes are a weighted average of all the observed covariates where each covariate is weighted in its importance in determined the outcome, (b) with several covariates selection individual covariates may working different directions making interpretation difficult and also, (c) with multiple covariates some point estimates may be statistically significantly different from zero by random chance.

historically. With positive assortative matching, the coefficient on teacher value-added would be positive and with negative assortative matching, it will be negative. If there is little systematic sorting of students to teachers the coefficient on pre-sample teacher value-added will be zero.

I present the results in the top panel of Table 5. In models with school and year effects in the second stage only the point estimates are very small, and none is statistically significant at the 5 percent level. This is evidence of minimal bias due to sorting *into* tracks. While the results thus far suggest that there is little selection to tracks, the main specification is based on variation *within* tracks. As such, it is important to show that there is no selection to teachers *within* tracks. I implement this additional test by augmenting the test described in IV.1 to include track-by-school fixed effects in the second stage as opposed to only school fixed effects. These estimates are presented in the lower panel of Table 5. As before, the point estimates are very small, and none of the coefficients is statistically significant at the 5 percent level. This is evidence of minimal bias due to sorting *within* tracks.

V Results

I present the estimated variance of the persistent teacher effects based on correlations across teachers' classrooms over time in Table 6. I present estimates from a variety of models that include different covariates and control for different sources of confounding variation. For each model I report the estimated standard deviation (based on the square root of the estimated covariance) of the teacher effects, the estimated 95% confidence upper and lower bounds of the estimated standard deviation, and also the *p*-value associated with the hypothesis that the correlations across classroom for the same teacher are equal to zero (i.e. the unbiased test).

In a basic model with one lag of achievement, student covariates and school fixed effects (Model 1) the estimated variance of algebra teacher effects on algebra test scores is 0.515σ (in student achievement units). Adding controls for the second lag of achievement (Model 2) reduces the estimate by 12 percent to 0.46σ. Including additional controls for mean classroom and school peer characteristics (Model 3) reduces the estimate slightly to 0.443σ. Model 4, which includes school-by-course-level effects (i.e. school*level of math course), student covariates, twice lagged achievement, and year effects yields an estimate of 0.417σ. Adding additional controls for mean classroom and school peer characteristics (Model 5) reduces the estimate slightly to 0.40σ. Accounting for school-by-track fixed effects has a sizable effect on the estimated effects. A model with school-by-track effects, student covariates, peer covariates and achievement, and twice-lagged achievement (Model 7) yields an estimate of 0.189σ. That is, including track-school effects reduces the estimated variance of teacher effects by more than half. Finally, in the preferred specification that also includes school-by-year effects (as opposed to separate school and year effects) the estimated variability of algebra teacher effects on algebra scores is 0.19σ. The null hypothesis of no algebra teacher effects is rejected at the 1 percent level and the 95 percent confidence interval for the sd of teacher effects is [0.108, 0.246].

The effects of English teachers on English scores are smaller than those for math. Specifically, in the basic model that includes one lag of achievement, student covariates and school fixed effects (Model 1) the estimated variance of English teacher effects on English test scores is 0.313 standard deviations (in student achievement units). However, one fails to reject the null hypothesis of no English teacher effects at the 5 percent level. Adding additional controls for the second lag of achievement and peer characteristics reduces the estimate to 0.25σ. Even for this intermediate specification, the *p*-value associated with the null hypothesis of no English teacher effects in 0.16 so that one cannot reject that there are no English teacher effects. In the preferred model with school-by-track effects, student covariates, peer covariates and achievement, twice lagged achievement, and school-by-year effects the estimate is 0.13σ and the *p*-value of no effect is 0.188. While the point estimates suggests that there may be some persistent English teacher effects, the statistical tests indicate that this may be due to random chance. I investigate this further in section V.3 where I look at the out of sample predictions.

As a check on the estimated specifications, one can look at the effect of math teachers on English performance and vice-versa. While two of the specifications looking at English teacher effects on algebra scores yield covariance that are non-zero, the estimates are negative so cannot be due to some persistent English teacher effect. In the preferred model, the point estimate is negative and the 95% confidence interval spans zero. For all specifications, one cannot reject the null hypothesis that there are no systematic effects of Algebra teachers on English scores at the 5 percent level. This is a useful test for bias due to tracking, and confirms the previous tests indicating that that there is little bias due to sorting.

The finding of no cross-subject spillover effects runs counter to findings in Aaronson et al (2007) and Koedel (2009) who use the *F*-test and find effects of math teachers on English scores and *vice versa*. To determine whether the differences across studies are due to different data sets, or due to the biases of the *F*-test in the teacher effects context, I estimated Model 7 on algebra test scores with English teacher indicator variables and Model 7 on English test scores with algebra teacher indicator variables. For both subjects, as expected, the *F*-tests reject the null hypothesis of no-spillovers across teachers as found in other studies. This suggests that the finding of spillovers across subjects in high school may have been an artifact of the biased *F*-test rather than real spillover effects from math to English and *vice versa*. This illustrates the importance of using the appropriate test.

V.1 Is this reduction in the variance of teacher effects due to over-controlling?

Because researchers typically find meaningful teacher effects in all subjects, readers may wonder if part of the reason that there are no test score effects for English is that there is little variation in teacher quality within school-tracks. There are few reasons why this is unlikely to explain the results. First, the very same procedure that finds no teacher effects for English does find treatment effects for Algebra. Second, because tracking is more common in math than in English, to the extent that teacher sorting to tracks reduces variation in teacher quality within tracks, this would drive the algebra effects towards zero rather than the English effects.

To assess the degree to which there is little variation in teacher quality within tracks, I estimate teacher value-added based on models without track-by-school fixed effects and then assess the degree to which there is variation within tracks in this measure of teacher quality. If the appropriate model is one in which there are no track-by-school fixed effects and the reason for the lack of English teacher effects is a lack of variation in teacher quality within tracks, then one should see that most of the variation in teacher value-added (estimated without track-by-school effects) occurs across tracks rather than within tracks. Table 7 shows the standard deviation of estimated teacher quality overall, within tracks, and across tracks. I also compute the fraction of the variance in estimated teacher value-added that occurs within tracks. For both algebra and English, approximately two thirds of the variance in teacher quality occurs within tracks so that if there is meaningful variation in teacher quality overall, there should also be meaningful variation in teacher quality within tracks. In fact, if the reduction in estimated variability is due to a reduction in the variance of true teacher quality within tracks of about 33 percent, then the estimated variances should be roughly 33 percent smaller when track-by-school effects are included. The results in Table 6 show that the estimated covariance for the algebra and English teacher effects fall by roughly 50 percent. As such, differences in variability within tracks alone cannot explain the difference in estimates that condition on track-by-school effects so that the difference is likely due to track-specific treatments within schools.

V.2 Is the lack of an effect for English due to measurement error?

Because the linking between teachers and student is not perfect in these data, there is the worry that the lack of an English teacher effect is due to certain teachers being linked to the wrong students. Such "measurement error" would lead one to understate the extent to which teacher effects are persistent over time. This is unlikely to drive the results because I do find strong persistent effects in algebra, which is subject to the same error. However, I test for this possibility parametrically, by seeing if the covariance of effects across classrooms is higher in schools where the scope for measurement error is zero. Specifically, if there is only one English I teacher at the school, and there is an English I teacher administering the test, we know that she is the correct teacher. If the lack of persistent English teacher effects were due to miss-measurement, then the covariance should be lower in schools with high chances of miss-measurement. I test this directly by regressing mean residuals for teacher *j* at time *t* on the mean residuals for teacher *j* at time *t-1* interacted with whether there is only one teacher (4 percent of the sample). If the covariance were not different for those schools where there is no miss-measurement, it would indicate that this is not the reason for the lack of a persistent English teacher effect. In fact, one fails to reject the null hypothesis that the covariance is higher where the scope for error is higher at the 10 percent level for both Algebra and English — so that measurement error does not explain the lack of persistent English teacher effects.

V.3 How predictive is estimated value-added of future teacher performance?

To verify that the estimated value-added has predictive power and to gauge the extent to which value-added estimates can predict a teachers performance in the future, I estimate teacher value-added using data from 2005 through 2007 and then use these estimates to see the effect on student test scores of these same teachers in the years 2008 through 2010. Specifically, I estimate equation [5] using 2005 through 2007 data, compute Empirical Baye's estimates of teacher value added based on mean teacher-level residuals and then estimate equation [6] below on the data from 2008 through 2010 where $\hat{\theta}_j$ is the estimated (pre sample) value added of teacher *j*.

$$[6] \qquad Y_{ijcy} = A_{iy-1}\delta + \psi\hat{\theta}_j + X_{iy}\beta + X^*_c \pi_c + I_{ci}\theta_{\mathrm{c}} + \theta_{\mathrm{sy}} + \varepsilon^*_{ijcy}$$

All variables are defined as before. I present results using both raw estimated value-added and Empirical Baye's value-added. If all of the estimated teacher effect is persistent over time, and there is no estimation error, then the coefficient on the raw value-added should be 1. This would indicate that a teacher who raises test score by 0.1σ in year *t* will also increase test score by 0.1 σ in year *t+1*). However, with estimation or measurement error the coefficient could be less than 1. With appropriate shrinkage estimates that account for estimation error, if all of the estimated teacher effect is persistent over time, then the coefficient on the raw value-added should be 1.

The results of this test are presented in Table 8. The estimate in specification 1 shows that an algebra teacher who raises algebra test scores by 1σ in the pre-sample raises algebra test scores by 0.0835σ in the analytic sample. This estimate is statistically significant at the 1 percent level. The fact that the coefficient on the raw value-added is much less than 1 either indicates that estimated teacher value-added is not persistent over time, or it is measured with substantial error. Specification 4 shows results based on the Empirical Baye's estimates that are shrunk towards zero to account for estimation error. This yields a statistically significant coefficient of 0.616, which indicates that after accounting for estimation error, an algebra teacher that would be expected to raise algebra test scores by 1σ in the pre-sample raises algebra test scores by 0.61σ in the analytic sample. This is consistent with about 40 percent of estimated value-added being non-persistent (Kane & Staiger, 2008).

These results suggest that while algebra teacher value-added persists over time, they are estimated with substantial error such that predicting future performance based on past performance in high school is limited. To illustrate this point consider the following calculation. The standard deviation of the Empirical Baye's estimate is 0.07σ and that of raw VA is 0.6σ so that moving from a median teacher to one at the 85th percentile of pre sample value added would improve test scores by between 0.04σ and 0.05σ. Similar calculations using elementary school data are much larger and range between 0.13σ and 0.16σ (Jackson and Bruegmann 2009).

Looking to English teacher value-added, the estimate in specification 7 shows that an English teacher who raises student English test scores by 1σ in the pre-sample raises algebra test scores by 0.0007σ in the analytic sample. This estimate is very small and is not statistically significant at traditional levels. Results from Empirical Baye's estimates also indicate that English teacher effectiveness estimated in the pre-sample is unrelated to teacher performance in the analytical sample.

As a further check on the main specification, I test for cross subject effects. The results in specification 7, 10, 13, and 16 all indicate that having an Algebra teacher with high pre-sample value-added has no effect on English scores in the analytic sample and having an English teacher with high pre-sample value-added has no effect on Algebra scores in the analytic sample. Again, these findings indicate that there are not cross-subject spillover effects and suggest that the finding of spillovers across subject was due to the use of the biased *F*-test.

In sum, the results indicate that high school teacher value-added has a much smaller signal to noise ratio than those in elementary schools. The predictive power of algebra value added in high school are about one-third of that for elementary math teachers and the predictive power for high school English teachers is zero. This suggests that using value-added in retaining high school teachers may have small effects in Algebra and no effect in English. This is an important practical difference from the results from the elementary school teacher literature.

V.4 Are these Effects Driven by Selection Bias Due to Sorting?

Even though the tests for sorting and selection to teachers suggest that sorting bias is minimal, readers may still worry that the effects are driven by non-random sorting within tracks in unobserved dimensions in a manner that the empirical tests conducted thus far cannot detect. The worry is that in any given year in a particular track the students who are better in some unobserved dimension are systematically assigned to a particular teacher while the weaker students in the cohort in the same track are assigned to another teacher. To test for selection within school-track-years in *unobserved* dimensions, I exploit the statistical fact that any selection within school-track-years will be eliminated by aggregating the treatment to the school-track-year level (leaving only variation within school-tracks but across years). Specifically, if the out of sample estimates obtained using variation in estimated teacher quality within school-track-years is similar to that obtained using variation across school-track years, then it would suggest that the estimates are not driven by selection to teachers within tracks in the same year. To test for this I estimate equation [7] and [8] below separately on the data from 2007 through 2010 where $\hat{\theta}_j$ is the estimated (pre sample) value added of teacher j, $\overline{\hat{\theta}}_{cy}$ is the mean estimated teacher value added in school-track c in year y, and θ_{cy} is a school-track-year fixed effect.[13]

$$[7] \qquad Y_{ijcy} = A_{iy-1}\delta + \psi_1\hat{\theta}_j + X_{iy}\beta + X*_c\,\pi_c + I_{cyi}\,\theta_{\text{cy}} + \theta_{\text{sy}} + \varepsilon*_{ijcy}$$

[13] Note that because this model is based on changes in mean value-added within a track over time, including estimated value-added for the year 2007 does not lead to any endogeneity (as the first year of data is based on changes in teacher quality within tracks between 2007 and 2008).

[8] $$Y_{ijcy} = A_{iy-1}\delta + \psi_2 \bar{\hat{\theta}}_{cy} + X_{iy}\beta + X^*{}_c \pi_c + I_{ci}\,\theta_{\mathrm{cy}} + \theta_{\mathrm{sy}} + \varepsilon^*{}_{ijcy}$$

By defining the treatment at the school-track-year level in [8], one is no longer comparing students within the same school-track-year, but only comparing students in the same school-track across different years where selection is unlikely. Conditional on track-school fixed effects, all the variation in this measure in [8] occurs due to changes in the identities of teachers in the track over time. If there is no sorting in unobserved dimensions, ψ_1 from [7] should be equal to ψ_2 from [8]. However, if all of the estimated effects are driven by sorting within school-track-years then there should be no effect on average associated with changes in mean teacher value-added in the track so that ψ_2 from [8] will be equal to *0*.

The results of this test are presented in Table 8. Because there are only effects for Algebra teachers, I do not discuss the results for English teachers (they are reported in the table). Columns 2 and 3 show the effect of pre-sample raw algebra teacher value-added on the analytic sample algebra test score using only variation within school-track-year cell and across school-track-year cells but within school tracks, respectively. The coefficients are 0.0871 and 0.0878 based on within and across variation, respectively. Using EB estimates in columns 5 and 6 yields coefficients of 0.604 and 0.564 based on within and across variation, respectively. For both sources of variation, the estimated effects are statistically indistinguishable from each other suggesting that the results are not driven by sorting to teachers in unobserved dimensions.

VI Conclusions

Even though there is mounting evidence that elementary school teachers have large effects on student test scores, much less is known about the effect of high school teachers on student outcomes. In this study, I argue that in a high school setting, even with random assignment to teachers, if different teachers teach in different tracks and students in different tracks are exposed to different treatments, there will be bias due to "track treatment effects". This additional source of bias creates new challenges to identifying teacher effects in high school.

I present an identification strategy that exploits detailed transcript data to allow for the credible identification of teacher quality effects in a high school setting. I show that using methods that account for "track treatment effects" yield estimated teacher effects that are substantially smaller than those obtained when track treatment effects are not accounted for. I find that a one standard deviation increase in persistent algebra teacher quality increases student achievement by about 0.19 standard deviations and find little evidence of any persistent English teacher quality effects on English test scores.

I demonstrate that the common practice of using the *F*-test on teacher indicator variables to test for the existence of teacher effects is problematic in the presence of classroom-level disturbances, and I present a new unbiased test that can be used. With this test, I find little evidence of spillovers across subjects even though the traditional *F*-tests indicate otherwise. This finding is important as it suggests that spillovers across subjects may not pose a major problem for identifying individual teacher effects in high school (as some researchers have suggested).

In sum, the results indicate that one should avoid the temptation to use studies based on elementary school teachers to make inferences about teachers in high school. The findings underscore the importance of using empirical methodologies that are appropriate to the specific high school context. From a policy perspective, this paper demonstrates that because of the smaller signal to noise ratio of value-added estimates for math teachers in high school, the potential gains of using value-added in personnel decisions in high school may be small. In addition, because English teachers have no effect on test scores, there may be important additional challenges in using test-score based measures of quality in the hiring and firing of English teachers and perhaps also for teachers in other subjects.

References

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

Chetty, R., Friedman, J., & Rockoff, J. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. *unpublished manuscript* .

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *Journal of Human Resources , 43* (5).

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). Identifying Effective Teachers Using Performance on the Job. *Hamilton Project Discussion Paper 2006-01* .

Hanushek, E. A. (2009). Teacher Deselection. In D. G. Hannaway, *Creating a New Teaching Profession* (pp. 165-180). Washington, DC: Urban Institute Press.

Jackson, C. K. (2010). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *NBER Working Paper 15990* .

Jackson, C. K. (2011). School Competition and Teacher Quality: Evidence from Charter School Entry in North Carolina. *NBER Working Paper No 17225* .

Jackson, C. K. (2009). Student Demographics, Teacher Sorting, and Teacher Qualty: Evidence From the End of School Desegregation. *Journal of Labor Economics , 27* (2), 213-256.

Jackson, C. K., & Bruegmann, E. (2009). Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics , 1* (4), 85-108.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics , 26* (1), 101–36.

Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER working paper 14607* .

Koedel, C. (2008). An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Department of Economics, University of Missouri Working Paper 0808* .

Koedel, C. (2008). Teacher Quality and Dropout Outcomes in a Large, Urban School District. *Journal of Urban Economics , 64* (3), 560-572.

Koedel, C., & Betts, J. (2009). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Working Papers 0902, Department of Economics, University of Missouri* .

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

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review , 94* (2), 247-52.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics* .

Figure 1: *F-tests without clustering (no teacher effects)*

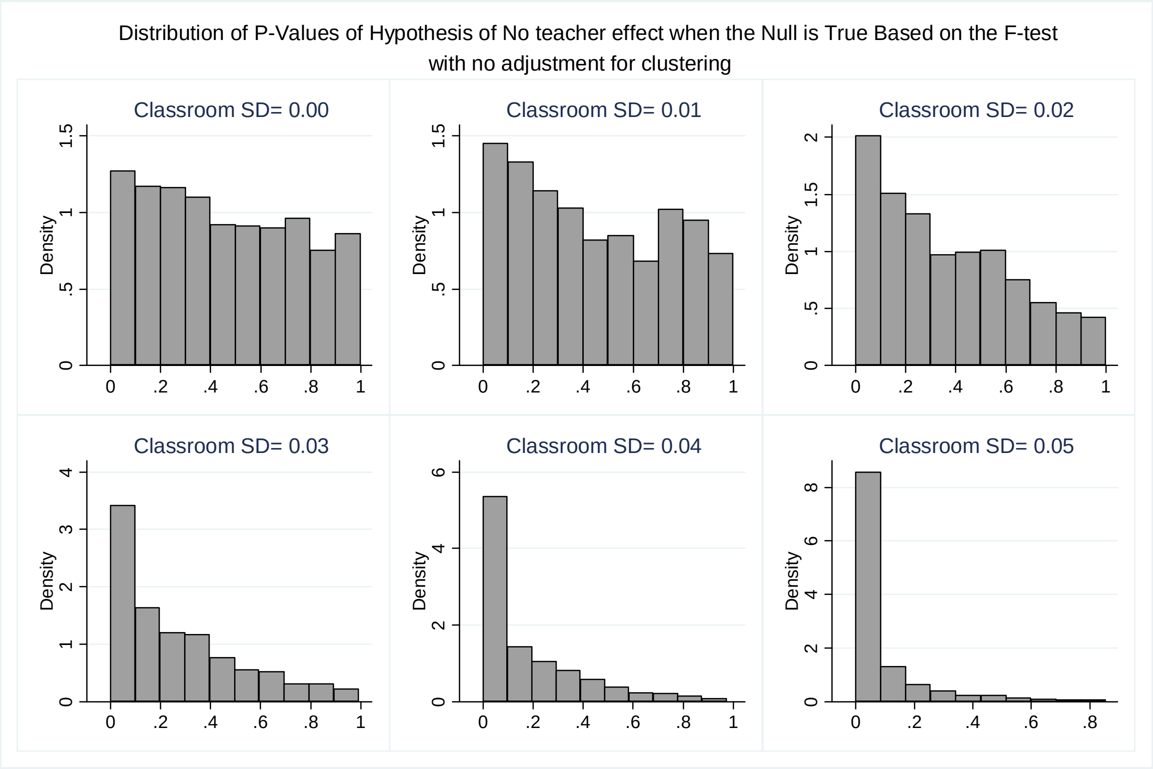

Figure 2: *F-tests with clustering (no classroom effects and no teacher effects)*

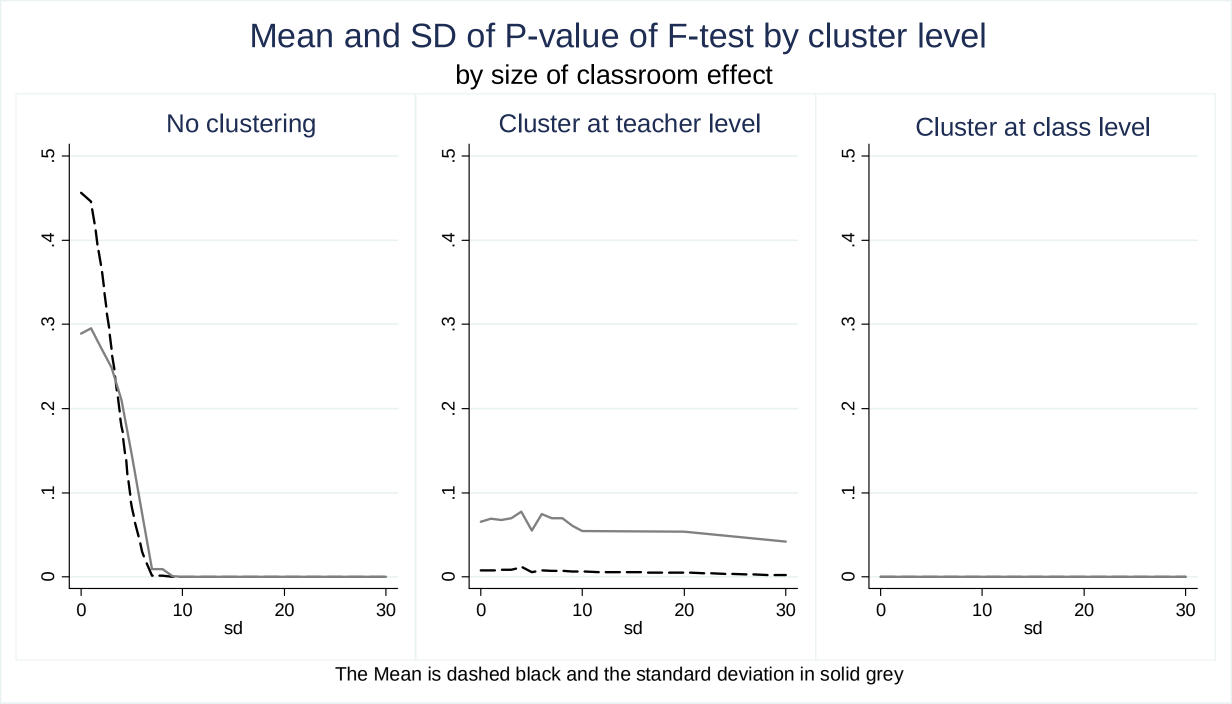

Figure 3: *Covariance Test under the presence of classroom errors (no teacher effects)*

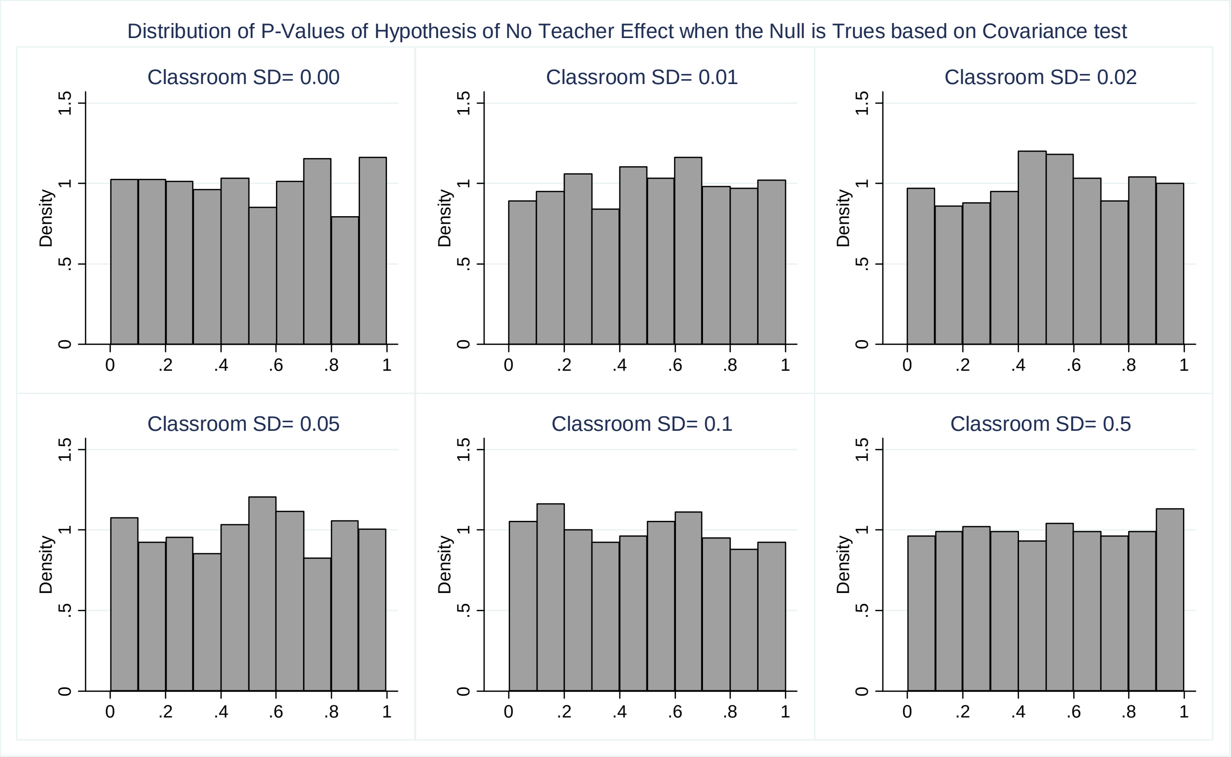

Table 1: *Summary Statistics of Student Data*

	Student level	School by Year Data	Track by School by Year
English I z-Score (9th grade)	0.039	-0.137	-0.202
	(0.980)	(0.895)	(1.169)
Algebra I z-Score (9th grade)	0.040	-0.096	-0.108
	(0.978)	(0.783)	(1.243)
Male	0.495	0.518	0.537
	(0.50)	(0.183)	(0.401)
Black	0.286	0.336	0.358
	(0.452)	(0.292)	(0.420)
Asian	0.020	0.014	0.016
	(0.020)	(0.013)	(0.095)
Hispanic	0.061	0.064	0.076
	(0.239)	(0.099)	(0.221)
Native American	0.020	0.016	0.011
	(0.106)	(0.080)	(0.091)
Math z-score 8th grade	0.0942	-0.156	-0.212
	(0.952)	(0.588)	(0.956)
Reading z-score 8th grade	0.0912	-0.141	-0.179
	(0.937)	(0.643)	(0.966)
Math z-score 7th grade	0.0873	-0.134	-0.166
	(0.949)	(0.646)	(0.972)
Reading z-score 7th grade	0.0917	-0.086	-0.151
	(0.911)	(0.663)	(0.922)
White	0.599	0.545	0.512
	(0.490)	(0.307)	(0.436)
Less than High School	0.038	0.047	0.043
	(0.193)	(0.089)	(0.192)
High School Graduate	0.277	0.273	0.292
	(0.447)	(0.277)	(0.396)
Trade school Graduate	0.027	0.025	0.024
	(0.163)	(0.057)	(0.121)
Junior College graduate	0.058	0.050	0.050
	(0.235)	(0.084)	(0.173)
Four-year college graduate	0.152	0.103	0.110
	(0.359)	(0.139)	(0.254)
Graduate school graduate	0.035	0.023	0.023
	(0.183)	(0.056)	(0.112)
Parental Education Missing	0.412	0.477	0.448
	(0.492)	(0.458)	(0.478)
Limited English Proficient	0.108	0.126	0.165
	(0.270)	(0.144)	(0.293)
Observations	377662	2499	58090

Notes:

Table 2: *Most Common Courses Taken*

course rank	Course Name	code	% of all students taking	Cumulative % of all courses taken
1	**English I***	1021	89.62	0.11
2	Physical Education	9011	84.25	0.22
3	**World History**	4024	83.82	0.33
4	**Earth Science**	3038	63.24	0.41
5	**Algebra I***	2023	50.91	0.47
6	**Geometry**	2030	19.93	0.50
7	Special Interest Topic	9520	19.00	0.52
8	Computer Applications I	6411	16.52	0.54
9	**Art I**	5415	15.94	0.56
10	**Pre Algebra I**	2020	15.22	0.58
11	**Biology I**	3020	14.51	0.60
12	**Intro to Algebra**	2018	14.00	0.62
13	**Basic Earth Science**	3040	13.21	0.63
14	**Spanish I**	1051	12.58	0.65
15	Digital Communications	6514	12.39	0.67
16	Special Interest Topic	1029	9.57	0.68
17	Team Sport	9015	9.23	0.69
18	ROTC	9501	9.09	0.70
19	Band	5255	8.38	0.71
20	Teen Living	7015	7.99	0.72

* Algebra I and English I can be taught at three levels (advanced, regular, and basic)

Table 3: *Illustration of the Variation*

		Track A	Track B	Track C	Track D	Track E
		Pre-Algebra	Alg I (reg)	Alg I (reg)	Alg I (reg)	Alg I (advanced)
		Eng I	Eng I	Eng I	Eng I	Eng I (advanced)
		Natural Sciences	Natural Sciences	Biology	Biology	Biology
		World History	US History	World History	US History	World History
				Geometry	Geometry	Geometry
					Spanish	Spanish
	Year					
Math Teacher 1	2000	X				
Math Teacher 2	2000		X	X		
Math Teacher 3	2000			X	X	
Math Teacher 4	2000		X			X
Math Teacher 1	2005	X				
Math Teacher 2	2005		X	X		
Math Teacher 3*	2005	-	-	-	-	-
Math Teacher 4*	2005	-	-	-	-	-
Math Teacher 5	2005		X			X
Math Teacher 6	2005			X	X	

Table 4: *Test for Sorting into Tracks by Prior Achievement*

	Math			Reading		
Average teacher-year level SD of variable	8th grade	7th grade	growth	8th grade	7th grade	growth
Actual	0.5750	0.5737	0.4990	0.6910	0.6676	0.5678
Full sorting within schools	0.1530	0.1683	0.1434	0.1720	0.1788	0.1502
Full sorting across schools	0.0012	0.0012	0.0012	0.0017	0.0010	0.0014
Random assignment within	0.6514	0.6360	0.4978	0.7445	0.7138	0.5645
Random assignment across	0.7411	0.7117	0.5106	0.7946	0.7593	0.5698

This table displays the average within-teacher-year standard deviation (i.e. the average across all teacher-years of the standard deviation of test scores computed for each teachers classroom in a given year) of 8th grade scores, 7th grade scores, and test-score growth between 7th to 8th grade for both math and reading. I present the actual within teacher-year test score dispersion, what one would observe with full student sorting (of the variable) within schools, full student sorting across all classroom and schools, random student assignment within schools and finally random student assignment across all classrooms and schools.

Table 5: *Further Evidence of no Selection Bias due to Sorting*

Dependent Variable	Predicted Algebra Score		Predicted English Score		
	Math Teacher	English Teacher	Math Teacher	English Teacher	
R^2 of Prediction	0.359	0.359	0.616	0.616	
Estimated VA	0.000622	0.0155	-0.00128	0.00659	Model 1
	[0.00669]	[0.0895]	[0.00589]	[0.104]	
Estimated VA	-0.00708	0.015	0.00134	0.0741	Model 2
	[0.00377]	[0.107]	[0.00398]	[0.0868]	

Model 1: School and year fixed effects

Model 2: School-by-Track fixed effects and year fixed effects

Robust standard errors in brackets clustered at the teacher level.

** p<0.01, * p<0.05

Table 6: *Estimated Standard Deviation of Teacher Effects*

Model	SD	Upper	Lower	p(t=0)	SD	Upper	Lower	p(t=0)
	English Teacher on Algebra Scores				Algebra Teacher on Algebra Scores			
1	-0.0667	0.148	-0.175	0.735	**0.515**	**0.585**	**0.433**	**0**
2	-0.0807	0.146	-0.185	0.639	**0.46**	**0.536**	**0.368**	**<0.001**
3	-0.0865	0.139	-0.186	0.578	**0.443**	**0.519**	**0.35**	**<0.001**
4	-0.0658	0.148	-0.174	0.74	**0.417**	**0.49**	**0.328**	**<0.001**
5	-0.0812	0.137	-0.178	0.602	**0.4**	**0.473**	**0.31**	**<0.001**
6	**-0.115**	**-0.0675**	**-0.147**	**0.002**	**0.189**	**0.253**	**0.0848**	**0.0122**
7	**-0.12**	**-0.077**	**-0.152**	**<0.001**	**0.177**	**0.242**	**0.0632**	**0.022**
8	-0.0806	0.0255	-0.117	0.0691	**0.19**	**0.246**	**0.108**	**0.0033**
	English Teacher on English Scores				Algebra Teacher on English Scores			
1	0.313	0.445	-0.048	0.0508	0.108	0.229	-0.171	0.57
2	0.264	0.401	-0.145	0.125	0.116	0.233	-0.165	0.508
3	0.245	0.383	-0.163	0.165	0.11	0.227	-0.166	0.542
4	0.261	0.395	-0.138	0.118	0.111	0.22	-0.153	0.491
5	0.243	0.377	-0.156	0.157	0.0986	0.21	-0.157	0.573
6	0.117	0.207	-0.125	0.354	-0.0903	0.0887	-0.155	0.309
7	0.101	0.194	-0.132	0.461	-0.098	0.0737	-0.157	0.202
8	0.134	0.212	-0.0961	0.188	-0.0539	0.105	-0.13	0.675

The covariance of mean classroom residuals across adjacent years within teachers are reported along with the upper and lower bound of the 95% confidence. The p-value associated with the null hypothesis that the estimated effect is less than or equal to zero is also reported.

Model 1: School effects, student covariates, lagged achievement, and year effects.
Model 2: School effects, student covariates, twice lagged achievement, and year effects.
Model 3: School effects, student covariates, peer covariates and achievement, twice lagged achievement, and
Model 4: School-by-course-level effects, student covariates, twice lagged achievement, and year effects.
Model 7: School-by-course-level effects, student covariates, peer covariates and achievement, twice lagged achievement, and year effects.
Model 6: School-by-Track effects, student covariates, twice lagged achievement, and year effects.
Model 7: School-by-Track effects, student covariates, peer covariates and achievement, twice lagged
Model 8: School-by-Track effects, student covariates, peer covariates and achievement, twice lagged achievement, and school-by-year effects.

Table 7: *Dispersion of Teacher effects (estimated without track fixed effects) across and within tracks*

	Math Teacher Effects				English Teacher Effects			
	SD raw	SD within Tracks	SD of track Means	% of Variance Within Tracks	SD raw	SD within Tracks	SD of track Means	% of Variance Within Tracks
Algebra scores	0.3572	0.2886	0.2101	**0.65**	0.0022	0.0018	0.0012	**0.67**
English Scores	0.0776	0.0643	0.0433	**0.69**	0.0405	0.0324	0.0242	**0.64**

Table 8: *Effect of Estimated Effects out of Sample at the Individual and Track level*

	Algebra I Score						English I Score					
	1	2	3	4	5	6	7	8	9	10	11	12
	Raw Value-Added			EB Value-Added			Raw Value-Added			EB Value-Added		
Own Algebra Teacher VA	0.0835*	0.0871*		0.616**	0.604**		0.000789	-0.0196		0.0159	0.0592	
	[0.0173]	[0.0199]		[0.123]	[0.141]		[0.0105]	[0.0104]		[0.0318	[0.0316	
Mean Algebra VA in track			0.0878*			0.564*			0.0384			-0.112
			[0.0202]			[0.148]			[0.0246			[0.0773]
School*Track FX	Y	inc	Y	Y	inc	Y	Y	inc	Y	Y	inc	Y
School*Track*Year FX	-	Y	-	-	Y	-	-	Y	-	-	Y	-
School*Year FX	Y	inc	Y	Y	inc	Y	Y	inc	Y	Y	inc	Y
Individual and peer controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	51,063	51,063	123,184	52,995	52,995	124,12	49,697	49,697	121,036	51,524	51,524	121,78

	13	14	15	15	16	18	19	20	21	22	23	24
	Raw Value-Added			EB Value-Added			Raw Value-Added			EB Value-Added		
Own English Teacher VA	0.00144	-0.00378		-0.0728	-0.0581		0.00619	0.0175*		-0.0885	-0.188	
	[0.0152]	[0.0146]		[0.0623	[0.0765		[0.00640	[0.00772		[0.0965	[0.158]	
Mean English VA in track			0.0715			-0.271			-0.016			0.0747
			[0.0429]			[0.154]			[0.0181			[0.226]
School*Track FX	Y	inc	Y	Y	inc	Y	Y	inc	Y	Y	inc	Y
School*Track*Year FX	-	Y	-	-	Y	-	-	Y	-	-	Y	-
School*Year FX	Y	inc	Y	Y	inc	Y	Y	inc	Y	Y	inc	Y
Individual and peer controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	45,510	45,510	119,193	50,179	50,179	121,36	82,565	82,565	201,848	87,677	87,677	204,32

Robust standard errors in brackets

** p<0.01, * p<0.05

All value-added covariates are estimated using pre-sample data from 2005 through 2007. The models with teacher value-added use data from 2008 through 2010. Models that are based on changes in value-added within a track use data from 2007 through 2010.

Appendix Note 1: *Empirical Bayes Estimates:* While teacher effects that come directly from [3] should yield consistent estimates of teacher value added, a more efficient estimate is the Empirical Bayes (EB) estimate that shrinks noisy value added estimates towards the mean of the value added distribution (in this case zero). Where u_j is random estimation error $\hat{\theta}_j = \theta_j + u_j$, and $\theta_j \sim N(0, Var(\theta))$, so that the total variance of the estimated effects is $Var(\hat{\theta}_j) = Var(\theta) + Var(u_j)$. With estimation error $E[\theta_j \mid \hat{\theta}_j] = (\sigma_\theta^2 / (\sigma_\theta^2 + \sigma_{u_j}^2)) \cdot \hat{\theta}_j$. The empirical analog of this conditional expectation is an EB estimate. I follow Kane and Staiger (2008) to compute the EB estimates. This approach accounts for the fact that teachers with larger classes will tend to have more precise estimates and there are classroom-level disturbances so that teachers with multiple classrooms will have more precise value added estimates. To simplify notation, I subsume all observed covariates into a single variable X_{ijy} , drop the track-school subscript c , and add a classroom-level error term υ_{jt} to re-write equation [3] as [A1] below.

$$Y_{ijy} = \eta X_{ijy} + \theta_j + \upsilon_{jy} + \varepsilon_{ijy} \qquad \text{[A1]}$$

In [A1], the total error term is $z_{ijt} = \theta_j + \upsilon_{jt} + \varepsilon_{ijt}$. Since the student error component is equal to zero in expectation, the mean residual for classroom *jy*, $g_{jy} = \theta_j + \upsilon_{jy}$, contains the teacher effect and the idiosyncratic classroom error. Since classroom errors are random, I use the covariance between mean residuals of adjacent classrooms for the same teacher $\text{cov}(g_{jy}, g_{jy-1}) = \hat{\sigma}_\theta^2$ as an estimate of the variance of true teacher quality. I use the variance of the classroom demeaned residuals as an estimate of $\hat{\sigma}_\varepsilon^2$. Since the variance of the residuals is equal to the sum of the variances of the true teacher effect, the classroom effect, and the student error, I compute the classroom error variance σ_v^2 by subtracting σ_ε^2 and σ_θ^2 from the total variance of the residuals.

For each teacher, I compute a weighted average of their mean classroom residuals, where classrooms with more students are more heavily weighted. Specifically I compute

$$\overline{\theta}_j = \sum_{t=1}^{T_j} c_{jy} \cdot \frac{1/[\sigma_v^2 + (\sigma_\varepsilon^2 / N_{jy})]}{\sum_{t=1}^{T_j} 1/[\sigma_v^2 + (\sigma_\varepsilon^2 / N_{jy})]} \quad . \qquad \text{[A2]}$$

Where N_{jy} is the number of students in class with teacher *j* in year *y*, and T_j is the total number of classrooms for teacher *j*. To obtain an EB estimate for each teacher, I multiply the weighted average of classroom residuals $\overline{\theta}_j$ by an estimate of its reliability. Specifically, I compute

$$\hat{\theta}_j^{EB} = \overline{\theta}_j \cdot \hat{\sigma}_\theta^2 / (\hat{\sigma}_\theta^2 + \sigma_{u_j}^2) \qquad \text{[A3]}$$

where $\sigma_{u_j}^2 = \left(\sum_{t=1}^{T_j} (1/[\sigma_v^2 + (\sigma_\varepsilon^2 / N_{jy})] \right)^{-1}$ is the estimation variance of the raw value added estimate. The shrinkage factor $\hat{\sigma}_\theta^2 / (\hat{\sigma}_\theta^2 + \sigma_{u_j}^2)$ is the ratio of signal variance to total variance and is a measure of how reliable an estimate $\overline{\theta}_j$ is for θ_j.