NBER WORKING PAPER SERIES

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, Revised September 2012
JEL No. H0,I20,J00

ABSTRACT

Unlike in elementary-schools, high-school teacher effects may be confounded with track-level treatments that are correlated with individual teachers. I document bias due to track-specific treatments, and show that traditional tests for the existence of teacher effects suffer from finite sample bias. Using new methods to account for these biases, I find modest algebra teacher effects and little evidence of English teacher effects. Unlike in elementary-school, value-added estimates of high-school teachers are weak predictors of teacher's future performance. The results indicate that teachers might not influence test-scores as much as previously thought.

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
SEPT 9, 2012
Northwestern University, IPR, and NBER

Unlike in elementary-schools, high-school teacher effects may be confounded with track-level treatments that are correlated with individual teachers. I document bias due to track-specific treatments, and show that traditional tests for the existence of teacher effects suffer from finite sample bias. Using new methods to account for these biases, I find modest algebra teacher effects and little evidence of English teacher effects. Unlike in elementary-school, value-added estimates of high-school teachers are weak predictors of teacher's future performance. The results indicate that teachers might not influence test-scores as much as previously thought. (JEL I21, J00).

There is consensus among policy-makers and researchers that teachers are the most important component of schools. This conclusion is based almost exclusively on studies of elementary-school teachers showing 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). This idea is further reinforced by recent findings indicating that effective elementary- and middle-school teachers are associated with improved long run outcomes (Chetty, Friedman, & Rockoff, 2011). In contrast, evidence on the importance of high-school teachers is sparse.

Using similar methodologies as those used for elementary school teachers, the few studies on high school teachers find similar results. Using value-added models, Aaronson, Barrow, and Sander (2007), Koedel (2008), and Goldhaber et. al. (2011) find that a standard deviation improvement in high school teacher quality raises student scores by between 0.10 and 0.4 standard deviations. Using variation across subjects for the same student, Slater, Davies, and Burgess (2011) find considerable variability in secondary school teacher effectiveness, and Clotfelter, Ladd, and Vigdor (2010) find that certain observable teacher characteristics are associated with differences in student test scores.

Because elementary-school students are typically exposed to one teacher and are all in the same academic track, while secondary-school students are exposed to several teachers and placed into different tracks, methodologies designed for elementary-school teachers may be inappropriate for measuring teacher quality in other grades. Specifically, in high-school settings, 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 other teachers (e.g. students who take Algebra I with Mr. Smith take physics with Mr. Black who has a direct effect on algebra scores), the content of other courses (e.g. students who take Algebra I with Mr. Smith also take physics which has a direct effect on algebra scores), or track-level treatments (e.g. college-bound students take Algebra I with Mr. Smith and are also part of a program that teaches study-skills that have a direct effect on algebra scores[1]). This source of bias has not been addressed in the few existing studies on high-school teachers, so it remains unclear whether, and to what extent, high-school teachers *really* affect student test scores.

In this paper, I aim to (a) demonstrate that the statistical methods used to identify teacher quality in elementary-school may be inappropriate for identifying teacher quality at the middle- or high-school level, (b) use data on high-school Algebra and English teachers in North Carolina to present evidence on the existence of track specific treatments, (c) employ a strategy to estimate the effects of algebra and English teachers on test scores in 9th grade that exploits detailed course-taking information to remove both selection bias across tracks *and* omitted variables bias due to track specific treatments, (d) present the first analysis of the extent to which historically estimated value-added predicts teacher effectiveness in a high-school setting, and (e) compare these results to those found for elementary-school teachers.

To address the unobserved track-level treatments problem, I estimate a standard value-added model with the addition of indicator variables for each academic track (the unique combination of school, set courses taken, level of courses taken). In such models, comparisons are made among students who are both 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 Algebra or English course *and in the same track* removes both the influence of any school-by-track level treatments that could confound comparisons of teachers, and 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 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

[1] For example, according to the mission statement, the Advancement Via Individual Determination (AVID) "program teaches the students how to study, read for content, take notes, and manage time."

estimate models that also 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 rather than to teachers *within* tracks, and I present several empirical tests that suggest little bias due to student selection.

While the focus of the paper is to estimate high-school teacher effects, this paper makes two methodological contributions to the broader teacher quality literature: The first contribution is to show that, under realistic conditions, the ubiquitous test for the existence of teacher effects is biased. It is common practice to use an *F*-test to determine whether there is excess heterogeneity in test scores across teachers beyond that explained by sampling variability. I show, via simulation, that with small transitory teacher-by-year level shocks (such shocks may arise due to shocks to teacher health, teacher effort, student health, shocks to classroom dynamics, or the proverbial "dog barking outside the exam room") the commonly used *F*-test will yield very small *p*-values even when there are no real teacher effects. This "over rejection" reflects finite sample bias because in most data we observe individual teachers for fewer than 30 years.[2] In the extreme case, where teachers are observed in only one year, one cannot distinguish a transitory teacher-year shock from a teacher effect (i.e., one cannot distinguish a good teacher from a lucky teacher). This is not solved by aggregating the data or clustering the standard errors.[3] To address this problem, I present an unbiased test for the existence of teacher effects that does not require a large number of teacher-year observations for each teacher, but only requires observing a large number of teachers in multiple years. This test tests whether, *on average*, teacher-by-year effects for the same teacher are correlated across years.

The second methodological contribution is to present an approach to estimating the variance of the variance of teacher quality effects. While existing studies present estimates of the variance of teacher quality effects, they do not present confidence intervals for these estimates.

[2] This is not enough years for the average of the teacher-by-year shocks to converge to zero for each teacher. Intuitively, with transitory teacher-by-year 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 transitory teacher-by-year effects for each teacher. In the extreme case, where teachers are observed in only one year, one cannot distinguish a transitory teacher-year shock from a teacher effect (e.g., one cannot distinguish a good teacher from a lucky teacher). While some researchers use the term classroom effect and teacher effect interchangeably, there is clearly a distinction between the two. While a teacher effect will certainly contribute to the classroom effect, the converse is not true. 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] This is a particular instance of the well-established fact that clustered standard errors can be downwardly biased when the number of clusters is small (MacKinnon and White 1985; Bertrand, Duflo, and Mullainathan 2004).

As such, it is unclear how much we should trust these estimates. Under certain conditions, the covariance of teacher-by-year effects for the same teacher across years is a consistent estimate of the variance of true teacher quality (Kane & Staiger, 2008). Because a covariance is a sample statistic, one can obtain confidence intervals for the variance of true teacher quality.

While I find little evidence of bias due to student sorting, I find evidence of bias due to omitted track-level treatments. After accounting for biases, a one standard deviation increase in Algebra and English teacher quality is *associated* with a 0.08 and 0.036 standard deviation increases in test scores, respectively. While biased F-tests reject the null of no teacher effects for both subjects, unbiased tests fail to reject the null of zero teacher effects for English. Consistent with this, a good Algebra teacher (85 percentile) based on previous performance raises current scores by 0.035 standard deviations more than an average teacher. In contrast, an English teacher's past performance is uninformative of her current performance. The results suggest that (a) the properties of high-school and elementary-school teacher value-added are meaningfully different, and (b) not all teachers affect student achievement. These findings have important implications regarding the use of test-score based measures of quality for high-school teachers.

Note that this paper speaks to the magnitude of persistent teacher quality. Several studies make this point that it is clearly incorrect to attribute all excess variation at the teacher-by-year level to teacher quality. As such, this paper takes the conservative approach and estimates the magnitude of persistent teacher quality (that one can be confident is due to the teacher). This approach is helpful because policies that advocate retaining/removing the best/worst performing teachers, are predicated on the notion that performance today is indicative of performance tomorrow. It is this predictable portion of teacher quality that is the focus of this paper.[4]

The remainder of the paper is as follows: Section II describes the data, Section III details the empirical framework, Section IV lays out 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. The student data

[4] Note that this is implicitly the approach taken in most papers on teacher quality and is the justification for the ubiquitous use of Empirical Bayes estimates that account for some sources of transitory variation.

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

The data cover 398,703 ninth grade students in 629 secondary-schools in classes with 4237 English I teachers, and 3559 Algebra I teachers. Roughly half of the students are male, 52 percent are white, 27 percent are black, 6.5 percent are Hispanic, and 1.7 percent is Asian. About 2.8 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, 23.6 with a high-school degree, 4.2 percent with a junior college or trade school degree, 11 percent with a four year college degree or greater (more than half of the data have missing parental education and are coded as such). About 11 percent of students are limited English proficient. All test scores are standardized to be mean zero with unit variance for each cohort and test.

Measuring Tracks:

Under tracking, different kinds of students (college bound, remedial, honors, etc.) take distinct groups of academic courses. In North Carolina, while there are hundreds of courses that students can take (these include special topics, physical education, and extracurricular activities) there are 10 core academic courses listed in Table 2 that make up over half of all courses taken. English I is the most common academic course (which 90 percent of 9th graders take), followed by World History, Earth Science, Algebra I, Geometry, Art, Biology I, Intro to Algebra, Basic Earth Science, and Spanish I. Because Algebra I and English I are taught at multiple levels (advanced, regular, and basic), students in the "high-ability" track will take these courses at the advanced level, while those in the "lower ability" tracks will take these courses at the basic level.

[5] The data link to student to the teacher who administered the test. In most cases this is the students own teacher but this is not always the case. I link classrooms in the testing files to classrooms in personnel files (with valid teacher identifiers). Classes that match across files on school, subject, year, class size, demographic composition, and teacher identifier, are considered perfect matches. See appendix note 1 for details on the matching procedure.

Because the advanced Algebra I class might emphasize different material from the regular class, even with random assignment of students to classes, basic comparisons of outcomes of Algebra I teachers will confound level of instruction effects with teacher quality. I can avoid this bias because the rich data contain information on the level of instruction.

I take as my measure of a school-track, the unique combination of the 10 largest academic courses, the 3 levels of algebra I, the 3 levels of English I, and the 629 high-schools. As such, all students who take the same set of courses in 9th grade, the same level of English I, the same level of Algebra I, *and attend 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 take either a different number of courses, or at least one different course, are in different school-tracks. In addition, students at the same school who take the same courses but took either Algebra I, or English I, at different levels are in different school-tracks. In principle, if students chose schools, courses, and course levels randomly, there would be millions of school-tracks. However, because many students pursue the same course of study, only 4 percent of students are in singleton tracks, over 60 percent are in school-tracks with more than 50 students, and the average student is in a school-track with 104 other students. Overall, there are 17,998 school-tracks with more than one student across the 629 high-schools. I present the standard deviations of the variables within-schools, and within-tracks-within-schools in Table 1, and discuss sorting into these tracks in section III.1.

III Empirical Framework:

To motivate the empirical strategy, I lay out a value-added model that includes both track-level treatments and transitory teacher-by-year level disturbances. I model the test score outcomes Y_{ijy} of student i with teacher j in school-track g in year y with [1] below.

[1] $$Y_{ijgy} = A_{iy-1}\delta + X_{iy}\beta + I_{jigy}\theta_j + \pi(P|g) + \mu_{jy} + \varepsilon_{ijgy}$$

Here, A_{iy-1} is incoming achievement level of student i, X_{iy} is a matrix of student-level covariates, I_{jigy} is an indicator variable equal to 1 if student i has teacher j in year y and equal to 0 otherwise, θ_j is a teacher fixed effect, $(P|g)$ is a treatment specific to students in school-track g, μ_{jy} is a random teacher-by-year level error such that $E[\mu_{jy}|J]=0$, and ε_{ijgy} is the idiosyncratic error term such that $E[\varepsilon_{ijgy} \mid J, X, A, G]=0$. When track-level treatments are unobserved and one does not account for teacher-by-year shocks errors, the OLS estimate of teacher effect $\hat{\theta}_j$ is given by [2].

$$[2] \qquad [\hat{\theta}_j \mid J, X, A] = \theta_j + \left(\sigma_{gj} / \sigma_j^2\right) \cdot E[\varepsilon_{ijgy} \mid g] + \left(\sigma_{gj} / \sigma_j^2\right) \cdot \pi(P \mid g) + \sum_{y=1}^{T_j} \frac{\mu_{jy}}{T_j} \mid J + \sum^{N_j} \frac{e_{ijgy}}{N_j}.$$

In [2], σ_{gj} is the covariance between teacher j and track g, σ^2_j is the variance of I_{jigy}, N_j is the number of students with teacher j, and T_j is the number of years observed for teacher j.

Potential Bias in Estimated Teacher Effects:

Equation [2] highlights that there are four distinct *potential* sources of bias in estimated teacher effects. The first potential source of bias is $A \equiv \left(\sigma_{gj} / \sigma_j^2\right) \cdot E[\varepsilon_{ijgy} \mid g]$. This is due to students selecting to tracks (and therefore teachers) based on unobserved characteristics that also affect test scores. For example, if highly motivated students select into the honors track, and Mr. Jones teaches students in the honors track, then students in Mr. Jones' class will on average be more motivated than students in other teachers' classes. This bias due to student selection has been discussed in Rothstein (2009), and Koedel and Betts (2012). The second potential source of bias is the last term $E \equiv \sum_{i=1}^{N_j} e_{ijgy} / N_j$. This term represents sampling variability. If each teacher is observed with more than 30 students (as is often the case), this terms will be small and will not result in bias. These two sources of bias have been discussed in existing studies. A third source of bias is the term $C \equiv \sum_{y=1}^{T_j} (\mu_{jy} / T_j) \mid J$, (i.e. the mean of the teacher-by-year level disturbances for a given teacher). This source of bias has been discussed extensively in Kane and Staiger (2008). This bias is due to transitory shocks that may be correlated with individual teachers *in finite sample*. For example, a teacher who teaches a class that had a flu outbreak will have lower performance than expected and, *based on only one year of results*, will appear to be less effective than her true ability. While this term should be near zero if teachers are observed in many years so that $T_j \rightarrow \infty$, because many teachers are rarely observed for more than 5 years, this bias term will be non-trivial and non-zero in short panels (i.e. virtually all existing datasets).

The potential source of bias not addressed by the existing literature is the term $B \equiv \left(\sigma_{gj} / \sigma_j^2\right) \cdot \pi(P \mid g)$. This bias is due to certain teachers being systematically associated with track-level treatments that directly influence test scores. For example, if Mr. Jones teaches algebra to students in the honors track, and honors classes teach students study skills that directly affect their algebra outcomes, one might erroneously attribute the benefits to the additional training in the honors track to Mr. Jones. This bias will exist even if students are randomly

assigned to teachers so long as teachers are correlated with track-specific treatments.

Section III.1 presents empirical evidence that bias due to sorting to tracks, track specific treatments, and transitory teacher-year shocks are likely to be non-trivial in high-school contexts, so that existing studies may overstate the importance of high school teachers.[6] Section IV presents a strategy to address these biases.

Transitory Shocks and Potential Bias in Test for the Existence of Teacher Effects:

Most studies on teacher quality estimate value-added models akin to equation [1] 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. Formally, researchers test $H_0 : \theta_1 = \theta_2 = \theta_3 = ... = \theta_J$ by comparing the computed *F*-statistic associated with this null hypothesis to an *F* distribution with N-J degrees of freedom (N is the number of observations and J is the number of estimated teacher effects). The F-statistic is computed as below where $\hat{\theta}$ is a J×1 vector of estimated fixed effects, $\bar{\hat{\theta}}$ is the mean of $\hat{\theta}_j$ s, and $\hat{V}_J$ is the J×J variance-covariance matrix for the teacher effects.

$$[3] \qquad F = \frac{1}{J-1}(\hat{\theta} - \bar{\hat{\theta}})'(\hat{V}_J)^{-1}(\hat{\theta} - \bar{\hat{\theta}}) \equiv \frac{\text{[accross-teacher variance]}}{\text{[within-teacher variance]}}.$$

The *F*-statistic is the ratio between the across- and within-teacher variance. If the variance of the estimated teacher fixed effects (across teacher variance) is large relative to the variance of the within-teacher variability in outcomes, this statistic will be large ─ implying that there is something systematic occurring at the teacher level that explains variability in outcomes.

With unaccounted for teacher-by-year shocks, the *F*-test is a biased test for the null hypothesis of no teacher effects and will tend to over-reject even when the null is true. One can think about this problem by considering the hypothesis that the *F*-test evaluates. The *F*-test tests for whether the means of the teacher-year-level outcomes for each teacher is the same across all teachers. To illustrate the importance of transitory teacher-by-year shocks, let us abstract away from any selection bias, track specific treatments, or sampling variability at the student level. In

[6] If one makes the common assumption that all errors in estimation are uncorrelated, then the variance of the raw estimated teacher effects from [2] is equal to the variance of the true teacher effects, plus the variance of any selection effects, plus the variance of any track level treatments, plus the variance of any teacher-year level errors, plus the estimation error. Formally, $Var(\hat{\theta}) = \sigma_\theta^2 + \sigma_A^2 + \sigma_B^2 + \sigma_C^2 + \sigma_D^2$. As such, $\sigma_\theta^2 = Var(\hat{\theta}) - \sigma_A^2 - \sigma_B^2 - \sigma_C^2 - \sigma_D^2$. It is common practice to estimate the variability of some noise component and then subtract this from the overall variance to obtain an estimate of true teacher quality. The downside of this approach is that failing to account for *all* sources of variability will lead one to overstate the importance of teachers. I argue that this has been the case for existing studies on high-school teachers.

this case [2] simplifies to [4] below.

[4] $$[\hat{\theta}_j \mid J, X, A] = \theta_j + \sum_{j=1}^{T_j} (\mu_{jy} / T_j).$$

One can see in [4] that the test of equality of the teacher-level mean outcomes for all teachers is really a test of equality of the teacher effects *plus the means of the transitory teacher-by-year shocks for each teacher*. This test is asymptotically equivalent to testing the equality of teacher effects only when each teacher is observed in several years so that $\sum\left(\mu_{jy} / T_j\right) \longrightarrow E[\mu_{jy} \mid J] = 0 \ \ \forall j$. However, if teachers are observed for a few years (as is usually the case), one cannot ignore the influence of transitory teacher-by year shocks, and the *F*-test will not be an appropriate test for the existence of teacher effects. The intuition for this problem is best illustrated by the extreme case where each teacher is observed in only one year. In this scenario, if there are large teacher-level mean residuals (due to large transitory teacher-year shocks), one might interpret this as evidence for the existence of teacher effects. However, in this simple case, the F-statistic will be large, not because there are teacher effects, but because it cannot distinguish between a good teacher and a lucky teacher.

To get a sense of the behavior of the commonly used *F*-test, I simulate hypothetical data with no teacher effects but in which there are teacher-by-year disturbances. I created a dataset with 200 teachers, each observed in 5 classrooms of 30 students each, 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 teacher-year-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 teacher-year-level disturbances of different sizes. If the test is unbiased, the *p*-values should follow a uniform distribution centered around 0.5.

With no teacher-year errors (SD=0.00) the distribution of *p*-values is roughly uniform. However, with just small teacher-year errors (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 modest teacher-year errors (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 estimated to be about as large as teacher effects themselves (Kane and Staiger 2008), the variance of the teacher-year effects are likely closer to 0.15σ. At this level, with no teacher effects, the *F*-test erroneously rejects the null more

than 99 percent of the time ─ rendering the extensively used *F*-test virtually uninformative.

The typical fixes for this kind of correlated shock would be to cluster the standard errors or aggregate the data to the level of the aggregate shock (Bertrand, Duflo, and Mullainathan 2004). Unfortunately neither of these solves the problem in this context. It is well documented standard adjustments for clustering are unreliable where the number of clusters is small. In this case, each teacher effect is based on only a handful of clusters (and sometimes only 1) so that clustering will not solve the problem.[7] Also, the- *F*-test will over-reject on aggregated data because aggregation does not deal with the problem that one cannot distinguish between a good teacher and a lucky teacher in short panels.[8] As such, much of the literature of teacher effects, which relies heavily on the *F*-test, might overstate the importance of teachers. In section IV, I present an unbiased test for the existence of teacher effects.

III.1 Evidence of Bias

Evidence of sorting of students into tracks

I present the standard deviations of the student variables within schools, and within tracks within schools in Table 1. Comparing the standard deviations within schools to those within school-tracks provides some indication of sorting into tracks. If there were no sorting to tracks within schools, these two estimated standard deviations would be the same. For all of student characteristics, there is greater variation within schools than there is within school-tracks. For example, the standard deviation of 8th grade math scores within schools is 0.894 while it is only 0.583 within school-tracks (a 55% reduction in the variance). Similarly, the standard deviation of having no parents with a high-school degree within schools is 0.413 while it is only 0.35 within school-tracks (a 27% reduction in the variance). Consistent with these tracks capturing important differences between college-bound, remedial, and average students, the standard deviation of the number of honors courses taken is 1.132 within schools while it is only 0.575 within school-tracks (a 74% reduction in the variance). In sum, the figures indicate that students are placed into tracks within schools in a manner that decreases dispersion in student characteristics within

[7] To illustrate this point, Figure A1 shows the mean *p*-value for different sizes of the teacher-by-year level errors. If the tests were unbiased, the *p*-values should be close to 0.5. Even with small shocks the mean *p*-values are much smaller than 0.5 when one clusters errors to account for dependence within classes or teachers.

[8] Moreover, aggregating the data means that each teacher effect is based on a very small number of observations. With a small number of observations per teacher, one cannot assume that each of the estimated effect comes from a normal distribution. Because the *F*-test requires that the each estimated coefficients be normally distributed, and asymptotic normality cannot be invoked with a small number of observations, the F-test will not be reliable.

tracks and groups students by ability.[9]

Evidence of sorting of teachers into tracks

To assess the degree to which track-specific treatments might be correlated with particular teachers, I computed the proportion of students in each teacher's class in each year who took at least on honors course (among the students' other courses). About 48 percent of all 9th grade students who took Algebra I or English I took at least one honors courses as one of their other courses. The standard deviation of the proportion of students in a teachers class in a given year in at least one honors class among other courses is 0.293 for Algebra I teachers and 0.32 for English I teachers. In other words, in both subjects roughly 20 percent of teacher-year observations have fewer than 25 percent of students who have taken at least one honors course while roughly 20 percent have more than 75 percent of students are in honors classes. Insofar as honors classes provide skills that affect Algebra I or English I scores directly, this would generate track specific treatment bias.

To assess the degree to which this varies systematically across teachers, I regress the proportion in at least one honors class among other courses for a given teacher in year *t* on the same proportion for the same teacher in year *t-1*. If there were no systematic relationship between teachers and tracks, the coefficient on the lagged outcome would be zero. The results (Table 3) show a strong association between the proportion in at least one honors class among other courses for a teacher in year *t* and year *t-1*. That is, within the same course (but not the same track), *some teachers consistently teach students who take more/less honors classes than the average student than other teachers*. This relationship holds across both subjects, and holds both across and within schools. This suggests that bias due to track-specific treatments might be a cause for concern.

Evidence of transitory teacher-by-year shocks

As discussed previously, transitory teacher-by-year level shocks can lead one to overstate the variance of teacher effects, misattribute transitory variation to individual teachers, and bias statistical tests for the existence of teacher effects. The extent to which this is a problem depends on the number of years each teacher is observed in the data and the size of the transitory teacher-by-year shocks. Aaronson, Barrow, and Sander use 2 years of data, while Koedel uses 1 year of data so that exiting studies do have enough observations to ignore any meaningful transitory

[9] In appendix note A3, I present evidence of modest student sorting to teachers based on prior achievement.

shocks. To gain a sense of the size of any transitory shocks, I first estimate a simple value-added model of student achievement with student covariates, school fixed effects, and year fixed effects. [10] Then, I estimate the variability of mean residuals within teachers across years, and compare this to the within teacher variability one would observe due to sampling variability alone. If the variance in the within teacher residuals is large relative to what would be expected due to sampling variability, it would imply that teacher-by-year level shocks exist and should be accounted for. I estimate the standard deviation of the sampling variability from year to year by computing the standard deviation of the student level residuals and dividing by the square root of the mean number of students in each year for each teacher (Kane and Staiger 2008; Aaronson Barrow and Sander 2004). The SD of the student level residuals is 0.54 and 0.58 for English and algebra, respectively. The average English teacher has 78 students per year and the average Algebra teacher has 58 students per year. As such, the implied mean SD of sampling variability for English I and Algebra I is 0.061 and 0.076, respectively. However, the SD of mean residuals within teachers across time is 0.277 and 0.30 English and algebra, respectively ─ much larger than what can be explained by sampling variability alone. This implies that teacher-by-year shocks exist, are sizable, and need to be taken into account.

IV Empirical Strategy

One can see that the variance of the estimated teacher effects $\hat{\theta}$ from [3] may overstate the variance of true persistent teacher quality because (a) this confounds teacher effects and track level treatments, and (b) teacher effects are estimated with error due to sampling variation and transitory teacher-by-year-level disturbances. I propose strategies to address this.

Removing bias due to track level treatments and selection

One can remove the influence of track-level treatments and selection to tracks by making comparison 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_{gi}, an indicator variable equal to 1 if student i is in school-track g and 0 otherwise. This leads to [5] below.

[5] $$Y_{ijcy} = A_{iy\text{-}1}\delta + X_{iy}\beta + I_{ji}\theta_j + I_{gi}\theta_g + \varepsilon_{igjy}$$

By conditioning on school-tracks, one can obtain consistent estimates of the teacher effects θ_j as

[10] I estimate $Y_{ijcy} = A_{iy\text{-}1}\delta + X_{iy}\beta + \theta_s + \theta_y + \varepsilon_{ijcy..}$

long as there is no selection to teachers *within* a school-track.[11]

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 4. There are two tracks A and B in a single school. There are two 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 and 2005 teacher 2 is replaced by teacher 3. Because, teachers 2 and 3 both teach in track B (in different years) one can estimate the value-added of teacher 2 relative to teacher 3 by comparing the outcomes of students in track B with teacher 2 in 2000 with those of students in tracks B with teacher 3 in 2005. 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 coincide with the hiring of new teachers), one can use the change in outcomes between 2000 and 2005 for teacher 1 (who is 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. I test for this explicitly in Appendix note 2, and find little evidence of selection to teachers on 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 1 and 2 taught students in track B in 2000 one can estimate the value-added of teacher 1 relative to that of teacher 2 by comparing the outcomes of teachers 1 and 2 among those students in track B in 2000. This source of variation is robust to student selection to tracks and is valid as long as

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

students do not select to individual teachers *within* tracks. In section V.2, I show that the findings are not driven by student selection within tracks.

To provide a sense of how much variation there is within tracks during the same year versus how much variation there is within tracks across years, I computed the number of teachers in each non singleton school-track-year-cell for both Algebra I and English I (Appendix Table A1). About 63 and 51 percent of all school-track-year cells include only one teacher in English and algebra, respectively. This implies that for more than half the data, the variation will be based on comparing single teachers across time within the same school track. However, 38 and 49 percent of school-track-year cells have more than one teacher for English and Algebra respectively, so that more than one-third of the variation is within tracks-year cells. In section V.4 I show that the results are similar when using each source of variation separately so that they might both be valid. Moreover, I show that the results are not driven by student selection.

Removing bias due to transitory teacher-year level shocks

While the variance of the estimated teacher effects $\hat{\theta}$ from [5] will not be biased due to track-level treatments, it will still overstate the variance of true teacher quality because of sampling variation and transitory shocks. I present two approaches to address this issue.

If one is willing to attribute any teacher-level variation in test scores not explained by teacher-year shocks or sampling variability to persistent teacher quality (a strong assumption), one can compute the variance of the raw teacher effects and subtract the estimated variance of the transitory teacher-year level shocks (Kane and Staiger 2005). I estimate the combined variance of the teacher-year level shocks and sampling variation with the variance of mean residuals from year to year *within teachers* (i.e. the within-teacher variability in average test scores over time). I then divide this by the number of year observations for each teacher to obtain an estimate of the variability in the raw teacher effects that can be attributed to transitory variability. Because this estimate attributes any variability not explicitly accounted for by teacher-year effects or sampling variability to teachers, this approach may overstate the true variance of persistent teacher quality. This motivates the approach below.

To estimate the variance of teacher quality I follow Kane and Staiger (2008) and Jackson (2010).[12] Specifically, I estimate equation [5] without teacher indicator variables, and take the covariance of mean teacher-by-year level residuals for the same teacher over time as my estimate

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

of the variance of the *persistent* component of teacher quality observed across years. Specifically, in the first stage I estimate equation [6] below.

[6] $$Y_{igjy} = A_{iy\text{-}1}\delta + X_i\beta + X^*_{jy}\pi + I_{gi}\,\theta_g + \theta_{sy} + \varepsilon^*_{ijgy}$$

The key conditioning variable is $I_{gi,}$ an indicator variable denoting the school-track g of student i. $A_{iy\text{-}1}$ 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 including parental education, ethnicity, gender, LEP status, number of honors courses taken, and whether the students has been designated gifted in English and/or math. I also include the mean incoming test scores *and characteristics* of other students with teacher j in year y X_{jy}*. 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}. The error term includes the teacher effect and the teacher-by-year effect so that $\varepsilon^*_{ijgy} = \theta_j + \mu_{jy} + \varepsilon_{ijgy}$.

In the second stage, I compute mean residuals from [6] for each teacher in each year $n_{jy}^{-1}\sum_{i=1}^{n_{jy}} e^*_{ijgy} \equiv \theta_j + \mu_{jy} + \bar{e}_{jgy}$, where n_{jy} is the number of students with teacher j in year y. To estimate the variance of the persistent teacher quality, I compute the covariance of mean residuals for the same teacher in years y and year $y\text{-}1$. If the non-persistent error components for each teacher e_{jgy} and μ_{jy} 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 $Cov(\theta_j, \bar{e}_{jgy}) = Cov(\theta_j, \bar{e}_{jgy-1}) = Cov(\bar{e}_{jgy}, \bar{e}_{jgy-1}) = Cov(\theta_j, \mu_{jy}) = Cov(\theta_j, \mu_{jy-1}) = Cov(\mu_{jy}, \mu_{jy-1}) = 0$ then $Cov(\varepsilon^*_{ijgy}, \varepsilon^*_{ijgy-1}) = \mathrm{var}(\theta_j)$.

Testing for Teacher quality Effects

Section III shows that the F-test on the estimated teacher effects is not a valid test for the existence of persistent teacher quality when there are transitory teacher-by-year-level shocks. Instead, I propose a test based on the simple idea that if teacher effects exist, the residuals for a teacher in one year should be correlated with residuals from another year 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 $t\text{-}1$ and then implementing a t-test for the coefficient on the mean lagged residuals. Even in the presence of large teacher-by-year disturbances, as long as teacher-by-year

disturbances are uncorrelated over time, this tests will be an unbiased test for the existence of persistent teacher quality effects. This test does not suffer from the finite sample problem of testing for equality of all the individual teacher effects because it only requires that one parameter be identified (i.e. the correlation between residuals for year t and year t-1) for all teachers on average rather than one parameter for each teacher. Identifying the correlation between residuals for year t and year t-1 requires only that there be enough teachers observed for more than one year. This new test does not require that each teacher effect be roughly correct (as is the case with the F-test), but only that any true teacher effects are persistent over time, and transitory shocks be uncorrelated over time for the same teacher.

To illustrate the unbiasedness of this proposed test, I show the covariance test's performance on the same simulated data where the *F*-test was problematic. Figure 2 shows the distribution of the *p*-values associated with the covariance tests across the 1000 replications for teacher-by-year disturbances of different sizes. If the test is unbiased, the *p*-values should follow a uniform distribution centered around 0.5. Irrespective of the size of the teacher-by-year errors, the *p*-values follow a uniform distribution, so that this test is robust to idiosyncratic teacher-by-year disturbances and will be an unbiased test for the existence of teacher quality effects.

Deriving confidence bounds for the variance of teacher quality effects

As discussed above, the covariance between mean teacher-level residuals in year *t* and year *t-1* is an estimate of the variance of true teacher quality. Because a sample covariance is a sample statistic, one can use a bootstrap to obtain consistent confidence intervals for the estimated covariance ─ and by implication the variance and standard deviation of true teacher effects. Such bounding exercises have not been done in the extant literature on teacher quality. From a policy perspective, this is important because it allows one to bound the kinds of improvements one can expect to see from using estimated value-added in retention and personnel decisions. Moreover, this is important because it allows one to determine the extent to which one can trust the exact point estimates from any particular study.

V Results

In Table 5, I present the estimated variance of effects under five models. Model 1: with both school and year fixed effects and lagged individual level achievement; Model 2: with school and year fixed effects and the rich set of all individual level covariates; Model 3: with school and

year fixed effects and all individual level and classroom peer level covariates; Model 4: with school-by-track and year fixed effects and all individual level and classroom peer level covariates; Model 5: with school-by-track fixed effects, school-by-year fixed effects and all individual level and classroom peer level covariates.

The estimated variability of raw Algebra teacher effects is similar across models that do not include school-track effects. For these three models the standard deviation of the raw teacher fixed effects is about 0.23σ (in student achievement units). These raw estimates are similar in magnitude to those in Aaronson Barrow and Sander (2007). In these models, the standard deviation of the mean residuals within teachers across years is about 0.35σ ─ indicating that there is more variation in teacher performance from year-to-year than there is across teachers. Because each teacher estimate is based on an average 3.1 years, this implies that approximately $0.35/\sqrt{3.1}=0.199\sigma$ can be attributed to transitory variation (i.e. due to teacher-by-year shocks and sampling variability). After accounting for transitory variability, *under the assumption that all the remaining variation reflects true teacher quality*, the implied standard deviation of teacher effects is approximately 0.12σ. In models that include school-track fixed effects and school-by-year effects (column 5) the standard deviation of the raw teacher fixed effects falls to 0.143σ. After adjusting for transitory variation, the implied variance of teacher effects falls to 0.08σ ─ suggesting that not accounting for track specific treatments or selection to tracks may lead one to overstate the importance of teachers by approximately fifty percent.

The results for English teachers reinforce the importance of accounting for potentially confounding track level disturbances. In models that do not include school-track fixed effects (columns 6,7, and 8), the standard deviation of the raw estimated teacher effects is about 0.16σ, and the adjusted estimate is about 0.06σ. In models that include track fixed effects and school-by-year effects (column 10) the standard deviation of the raw teacher fixed effects falls to 0.084σ. After adjusting these estimates for transitory variation, the implied standard deviation of teacher effects falls to 0.036σ. In fact, column 9 indicates that, in some specifications, after including school-track fixed effects, transitory variation can explain *all* of the variability in mean outcomes across English teachers.

As one might expect, the potentially biased F-tests of equality of the teacher level residuals is rejected for all models. However, the test for covariance of mean teacher-by-year level residuals across years for the same teachers supports a different conclusion. For Algebra, in

models that do not include school-track effects, one can reject that there is zero covariance in the teacher by year level residuals for the same teacher over time at the 1 percent level. However, in models that do include school-track effects, one cannot reject that there is zero covariance in the teacher by year level residuals for the same teacher over time at the 5 percent level, but can at the 10 percent level. This indicates that selection to tracks or track specific treatments may be important sources of covariance in teacher effects across years that should not be attributed to teacher effects. Models that include track fixed effects yield an estimated standard deviation of true teacher quality of approximately 0.11σ – similar to the adjusted estimates. The confidence interval reveals that this point estimate is not precisely estimated. The 95% percent confidence interval ranges from 0σ to 0.19σ. That is, the most credible set of estimates are sufficiency noisy that going from an average teacher to a teacher at the 85th percentile of the teaching quality distribution will increase Algebra test scores by anywhere between 0 and 0.17σ.

The results for English teachers differ substantively from those for Algebra teachers. For English, one fails to reject that there is zero covariance in the teacher by year level residuals for the same teacher over time at the 20 percent level in all models. In fact, for all models the estimated covariance is *negative* – which would imply no persistent teacher effects. This is consistent with the previous finding that all (or almost all) of the variability in mean teacher level residuals for English teachers can be explained by transitory variation. Models that include track fixed effects yield a 95% percent confidence interval for the standard deviation of between 0σ to 0.067σ. That is, while all the evidence suggests no English teacher effects, the estimates are sufficiency noisy that one cannot rule out English teacher effects smaller than 0.067σ with 95 percent confidence. The differences between the covariance-based estimates and the raw standard deviations underscore the fact that excess variability in test scores that can be attributed to individual teachers is not necessarily indicative of persistent differences in teacher quality.

V.2 Is the Lack of an English Teacher Effect a Statistical Artifact?

Because this study is the first to show that English teachers might not affect student test scores, it is important to establish that this lack of an effect is real, and is not a result of methodology or data limitations. As such, in this section I discuss, and rule-out, three statistical explanations for the lack of any English teacher effects.

Are there test ceiling or test floor effects that reduce the ability to detect English teacher effects?

If test scores suffered from ceiling/floor effects, then many students with different latent ability would have the same maximum/minimum English score, making it impossible to detect teacher effects for students in this ability range. One can assess this by looking for excess density of English test scores at the top or bottom of the distribution. I present kernel density plots for standardized English and Algebra test scores in appendix Figure A2. Both tests have single peaked bell-shaped distributions, and there is no visible evidence of any ceiling or floor effects.

Is the lack of an effect for English teachers due to measurement error?

Because the linking between teachers and students 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 persistent effects in algebra, which is subject to the same error. However, I test for this possibility by restricting the analysis to only those teachers who are perfectly matched[13]. For the subsample of English teachers for whom there is no "measurement error", the estimated covariance of mean residuals across years for the same teacher is small and negative ─ so that measurement error does not explain the lack of persistent English teacher effects.

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

Readers may wonder if part of the reason for no test score effects for English is that there is little variation in teacher quality within school-tracks. There are *a priori* reasons why this is unlikely to explain the results. First, the covariance between teacher-year level residuals within teachers across time is negative in all models (including those that do not include school-track effects). As such, the lack of persistent English teacher effects cannot be attributed to the inclusion of school-track effects. Second, the very same procedure that finds no teacher effects for English does find teacher effects for Algebra. To shed further light on this issue, I estimate teacher value-added based on models without track-by-school fixed effects (model 3 from Table 7) and then assess the degree to which there is variation across and within tracks in this measure of teacher quality. For both algebra and English, approximately two thirds of the variance in teacher quality (estimated without school-track effects) occurs within school tracks (Appendix Table A1) so that if there is meaningful variation in teacher quality overall, there is also

[13] These are either the only English I teacher in a school in a given year, or those teachers who are listed in the testing files and also have perfectly matching class characteristics. See appendix Note 1 for details.

meaningful variation in teacher quality within tracks.

Do Track Effects Remove the Effect of Teacher Experience?

Readers may wonder if conditioning on school tracks eliminates the benefits to teacher experience for high school teachers documented by Clotfelter et. al. (2010). To test this, I regress of English I and Algebra I scores on indicator variables for each year of teacher experience. I estimate models with no controls, with school fixed effects, and then with track-by-school fixed effects. I plot the estimated coefficients in Figure 3. Models with no controls or school fixed effects indicate that students have higher English I scores when they have teachers with more years of experience (top panel). However, conditional on school-track fixed effects, there is no systematic relationship between teacher experience and English scores. This is consistent with Table 5 that shows that, conditional on track fixed effects, English teachers do not affect test scores. In contrast, in all models, students perform better with Algebra teachers who have more years of experience (lower panel).

V.3 How predictive is estimated value-added of teachers' future 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 Bayes estimates of teacher value added based on mean teacher-level residuals, normalize the Empirical Bayes estimates[14] to be mean zero unit variance, and then estimate equation [7] below on the data from 2008 through 2010 where $z_{\hat{\theta}_j}$ is the estimated (pre sample) normalized value added of teacher j.

[7] $$Y_{ijcy} = A_{iy-1}\delta + \psi z_{\hat{\theta}_j} + X_{iy}\beta + X^*_{jy}\pi + I_{gi}\theta_g + \theta_{sy} + \varepsilon_{ijgy}$$

All variables are defined as before. The results are presented in Table 6.

Column 2 shows that a one standard deviation increase in estimated pre-sample value-added (going from a median teacher to a teacher at the 85th percentile) raises algebra test scores by 0.0238σ. This effect is statistically significant at the 1 percent level. The magnitude of this effect is noteworthy. A very similar exercise for elementary-school teachers in North Carolina

[14] See appendix note 3 for a detailed description of how the Empirical Bayes estimates are formed from the raw teacher-level residuals.

finds that a 1 standard deviation increase in estimated pre-sample value-added raises math scores by 0.17σ (Jackson & Bruegmann, 2009). This indicates that value-added estimates in high-school have about 14 percent of the the out of smaple predictive power than those at elemnetary school. To ensure that the lack of predictive power is not due to "over controlling", I estimate the same model where the first stage value-added estimates do not account for track effects (not shown). Value added estimates that do not account for school track fixed effects have *less* the out of sample predictive power─ suggesting that the differences in out-of-sample predictability are not driven by differences in methodology.

Looking to English teachers, the estimates in specification 7 show that a one standard deviation increase in estimated pre-sample value-added raises English test scores by 0.0124σ. This effect is statistically significant at the 1 percent level. The small magnitude of this effect is notable because this is about 14 percent of the size of out-of-sample predictive ability of estimated value-added for teachers at the elementary-school level.

It is important to note that if there is systematic selection to teachers within tracks based on *unobservable* student characteristics, then one might find that teachers with high estimated value-added in the past also have higher value-added in the current sample (even if there are no real teacher effects). In such a scenario, having high value-added in the pre-sample might indicate that certain teachers always teach those students who over-perform relative to observed characteristics rather than reflecting the real casual effect of the teacher. As such, it is important to obtain estimates of the out of sample predictive power of teacher value added that are robust to student sorting within school-tracks.

V.4 Are These Out of Sample Effects Driven by Student Sorting Within Tracks?

To test for student selection to teachers within school-track-years, 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 across years within school-tracks). 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, it would indicate that the estimates are not driven by selection to teachers within tracks. This is very similar in spirit to the test presented in Chetty et al (2011). To test for this, I estimate equation [8] and [9] below separately on the data from 2008 through 2010 where $z_{\hat{\theta}_j}$ is the normalized estimated (pre

sample) value added of teacher j, $\bar{z}_{\hat{\theta}_j}$ is the mean normalized estimated teacher value added in school-track g in year y, and θ_{gy} is a school-track-year fixed effect.

[8] $$Y_{ijcy} = A_{iy-1}\delta + \psi_1 z_{\hat{\theta}_j} + X_{iy}\beta + X^*_{jy}\pi + I_{gyi}\,\theta_{gy} + \theta_{sy} + \varepsilon^*_{ijcy}$$

[9] $$Y_{ijcy} = A_{iy-1}\delta + \psi_2 \bar{z}_{\hat{\theta}_j} + X_{iy}\beta + X^*_{jy}\pi + I_{gi}\,\theta_{g} + \theta_{sy} + \varepsilon^*_{ijcy}$$

In [8] because the model includes school-by-track-by-year effects and teacher quality is defined at the student level, the variation all comes from comparing students in the same school-track in the same year but who have different teachers ─ i.e. the variation that might be subject to selection bias. In contrast, by defining the treatment at the school-track-year level in [9], one is no longer comparing students within the same school-track-year, but only comparing students in the same school-track across different cohorts where selection is unlikely. Conditional on track-school fixed effects, all the variation in this aggregate teacher quality measure in [9] 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 [8] should be equal to ψ_2 from [9]. However, if all of the estimated effects are driven by sorting within school-track-years, there should be no effect on average associated with changes in mean teacher value-added in the track so that $\psi 2$ from [9] will be equal to 0.

The results of this test are presented in Table 6. For Algebra, using only variation within school-track years (column 3) yields a point estimate of 0.024, and using changes in aggregate track level mean teacher quality (using only variation across years) yields a point estimate of 0.0386 (column 4). This pattern is robust to including school by year effects ─ suggesting that the estimated algebra teacher effects are real and are not driven by selection to teachers within tracks. For English however, the point estimates within school-track years is 0.0138 (column 8) while those for models that are robust to selection are close to zero and not statistically significant (columns 9 and 10). This suggests that the small English effects estimated using within school-track-year variation might have been due to some mild selection on *unobservables*.[15] Taken at face value, the point estimate of 0.00034 indicates that policies that select high English value-added teachers will not yield meaningfully improved test scores.

VI Conclusions

Despite mounting evidence that elementary-school teachers have large effects on student

[15] In appendix note 2, I present weak evidence of selection on observables for English but not for Algebra teachers.

test scores, much less is known about the effect of high-school teachers. 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 also 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 teacher-by-year level disturbances, and I propose an unbiased statistical test.

I present an identification strategy that allows for the credible identification of high-school teacher effects. Using methods that account for "track treatment effects" yield estimated teacher effects that are considerably smaller than those obtained when track treatment effects are not accounted for. I find that a one standard deviation increase in algebra teacher quality is associated with increased student achievement of about 0.08 standard deviations. However, there is little evidence of any persistent English teacher quality effects on English test scores. I also show how using the F-test to test for the existence of English teacher effects leads to the wrong conclusion, and may have affected inferences in existing studies.

While the results indicate that there are real persistent Algebra teacher effects, the magnitude of the effects in this study are smaller than those found in other studies ─ indicating that accounting for transitory teach-by-year shocks and track specific treatments is important. Also, this is the first study to show that English teachers do not affect test scores ─ underscoring the importance of not generalizing teacher effects obtained in elementary school across all grade levels and subjects. It is important to note that the lack of any test score effects for English teachers does not necessarily mean that high-school English do not matter. It is possible that English teachers have meaningful effects on important non-cognitive outcomes that are not well-captured by test scores such as self-esteem, motivation, and aspirations.

I also investigate the extent to which estimated value-added predict a teacher's future performance. After accounting for biases due to student sorting, the results suggest that the scope for using value-added in personnel decisions in high-school might be limited. With a normal distribution, removing the bottom 5 percent of teachers increases average teacher quality by 0.1 standard deviations in value-added. The out of sample estimates suggests that this would raise student achievement by $0.0345*0.1=0.00345\sigma$ in Algebra (and no effect in English). A more aggressive policy of removing the lowest 30 percent of teachers would increase average teacher

quality by 0.5 standard deviations and would raise student achievement by 0.0345*0.5=0.017σ (1.7 percent of one standard deviation) in Algebra (with no effect on English performance). These calculations suggest that, even with large changes to the teacher quality distribution, one cannot expect very large improvements in student achievement associated with retaining the highest value-added teachers. This is because the ability for value-added to predict teacher performance out of sample is about 7 times smaller in high-school than in elementary school.

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 general. 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 the potential gains of using value-added in personnel decision in high-school may be small. In addition, because English teachers are found to have no effect on test-scores, using test-score based measures of quality in the hiring and firing of high-school English teachers may be misguided.

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.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119(1):249–75.

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).

Goldhaber, Dan, Peter Goldschmidt, Philip Sylling, and Fannie Tseng. *Teacher Value-Added at the High School Level: Different Models, Different Answers?* Working paper no. 2011-4.0. Center for Education Data & Research (CEDR), University of Washington, 2010.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). Identifying E¤ective 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* .

Mansfield, Richard K. (2011). "Teacher Quality and Student Inequality" Working paper.

MacKinnon, James and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29(3):305-325.

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* .

Slater, H., Davies, N. M. and Burgess, S. (2011), Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England. *Oxford Bulletin of Economics and Statistics*.

Figure 1: *F-tests without clustering (based on simulated data with no teacher effects)*

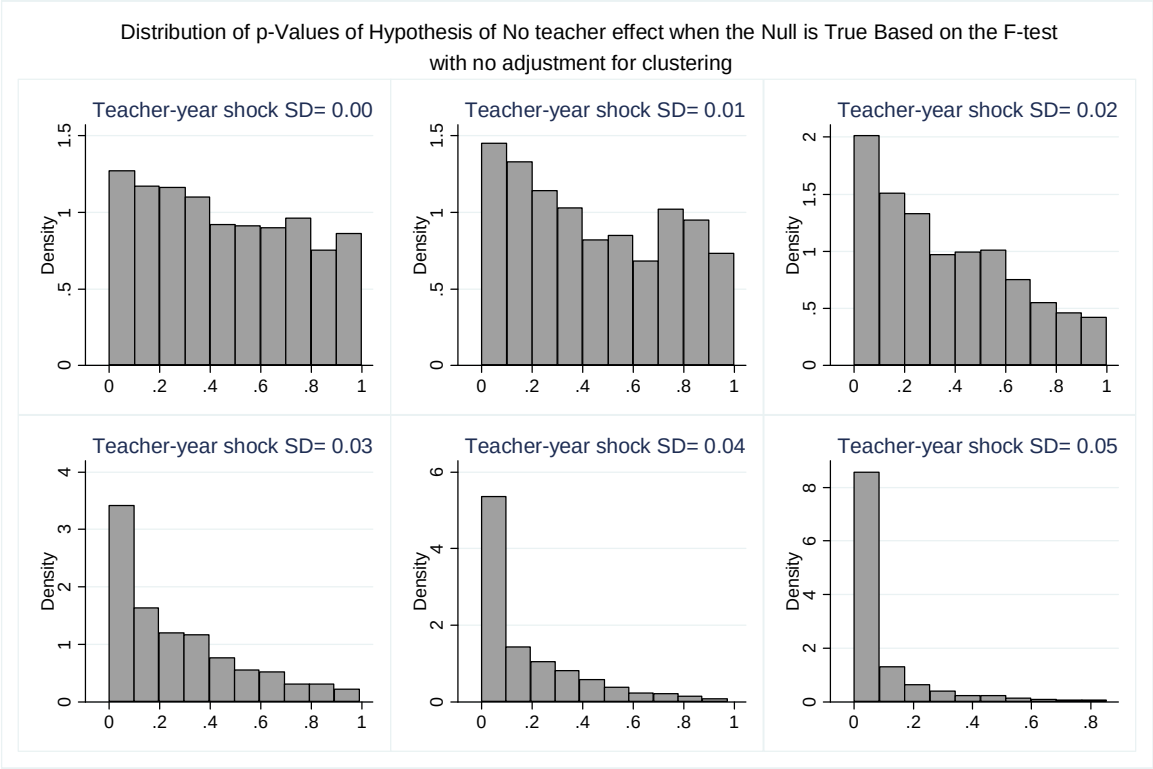


Figure 2: *Covariance Test (based on simulated data with no teacher effects)*

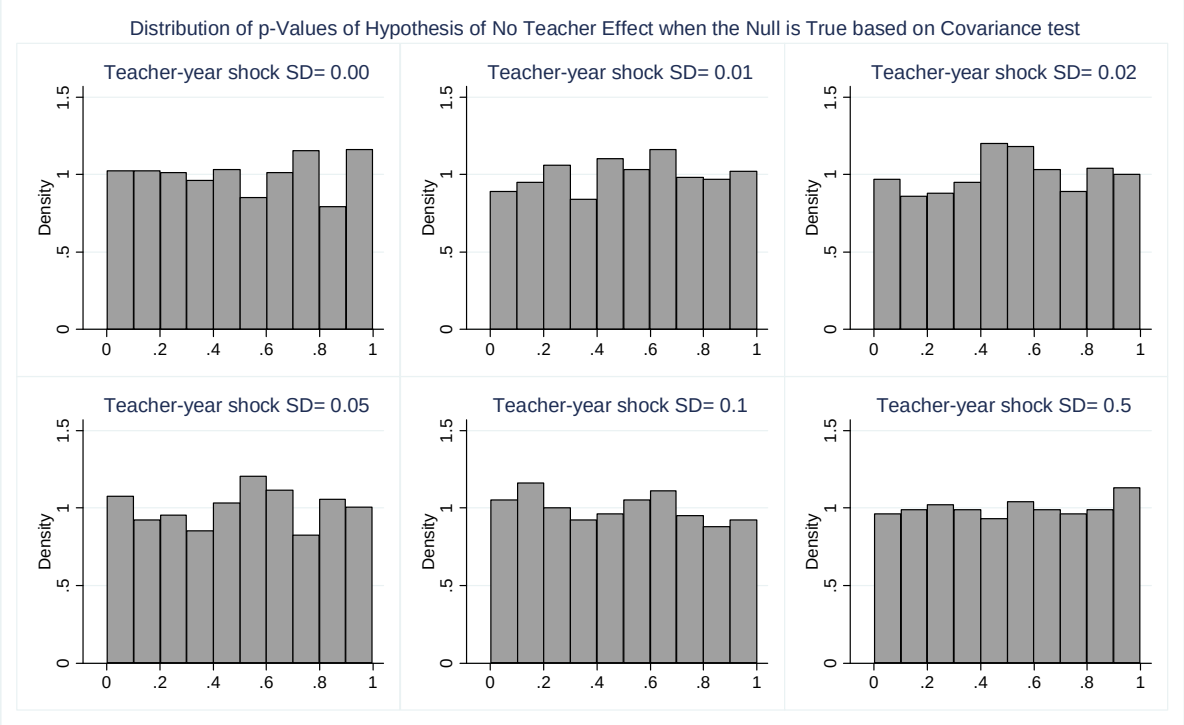
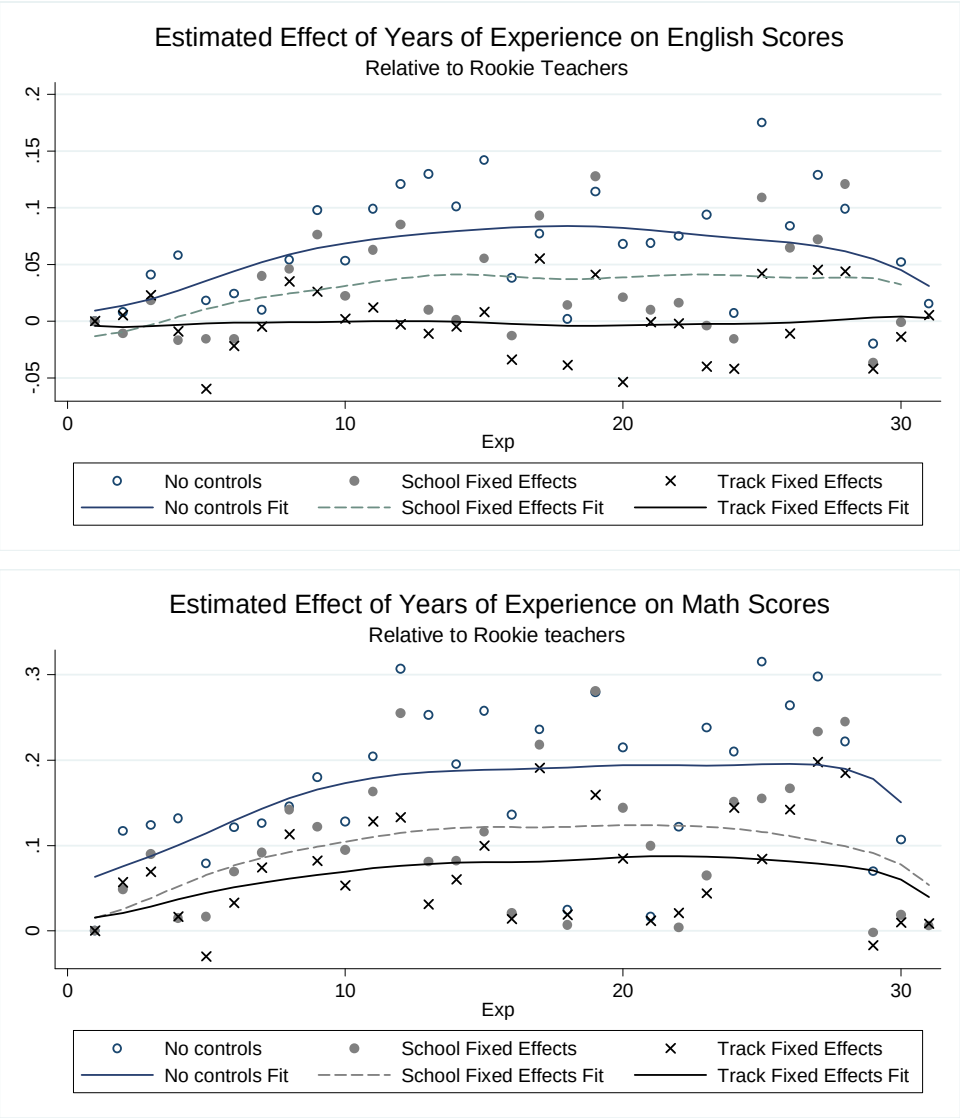


Figure 3: *The Marginal Effect of Teacher Experience under Different Models*



These figures present the estimated coefficients on indicator variables denoting each year of experience on English and Math test scores. The figures show the estimated point estimates for each year of experience and a lowess fit of the point estimates for models with no controls, school fixed effects, and track-by-school fixed effects.

Table 1: *Summary Statistics of Student Data*

Variable	Mean	Std. Dev.	Std. Dev. within school-tracks	Std. Dev. within schools
English I z-Score (9th grade)	-0.037	(1.001)	(0.766)	(0.937)
Algebra I z-Score (9th grade)	-0.052	(0.995)	(0.699)	(0.955)
Male	0.509	(0.500)	(0.470)	(0.498)
Black	0.270	(0.444)	(0.362)	(0.396)
Asian	0.017	(0.13)	(0.123)	(0.128)
Hispanic	0.065	(0.246)	(0.228)	(0.24)
White	0.527	(0.499)	(0.404)	(0.443)
Math z-score 8th grade	-0.025	(0.949)	(0.593)	(0.894)
Reading z-score 8th grade	-0.022	(0.948)	(0.613)	(0.908)
Math z-score 7th grade	-0.023	(0.951)	(0.673)	(0.902)
Reading z-score 7th grade	-0.014	(0.933)	(0.676)	(0.900)
Parental education: Less than High-school	0.028	(0.425)	(0.35)	(0.413)
Parental education: High-school Graduate	0.236	(0.161)	(0.135)	(0.161)
Parental education: Junior College graduate	0.042	(0.201)	(0.190)	(0.200)
Parental education: Four-year college graduate	0.108	(0.310)	(0.273)	(0.303)
Parental education: Graduate school graduate	0.023	(0.150)	(0.139)	(0.148)
Parental education: Missing	0.527	(0.499)	(0.384)	(0.481)
Limited English Proficient	0.111	(0.163)	(0.146)	(0.161)
Number of honors courses taken	0.775	(1.274)	(0.575)	(1.132)
Gifted in reading	0.017	(0.131)	(0.118)	(0.129)
Gifted in math	0.017	(0.130)	(0.118)	(0.128)
Honors student	0.386	(0.487)	(0.274)	(0.451)
Observations			398703	

Table 2: *Most common academic courses*

Academic course rank	Course Name	Course code	% of 9th graders taking	% of all courses taken
1	English I*	1021	90	0.11
2	World History	4024	84	0.11
3	Earth Science	3038	63	0.09
4	Algebra I*	2023	51	0.06
5	Geometry	2030	20	0.03
6	Art I	5415	16	0.03
7	Biology I	3020	15	0.02
8	Intro to Algebra	2018	14	0.02
9	Basic Earth Science	3040	13	0.01
10	Spanish I	1051	13	0.02

Table 3: *Distribution and persistence of honors students among teachers*

	1	2	3	4
	Algebra I		English I	
	Proportion of students in at least one Honors class among other courses		Proportion of students in at least one Honors class among other courses	
Lag of the proportion in at least one Honors class among other courses	**0.454** **[0.014]*** **	**0.225** **[0.016]*** **	**0.585** **[0.014]*** **	**0.449** **[0.017]*** **
Year Fixed Effects	Y	Y	Y	Y
School Fixed Effects	N	Y	N	Y
SD of mean outcome at teacher year level	0.293		0.32	
SD of mean outcome at teacher level	0.304		0.32	
Observations	12987		10855	

Standard errors in brackets
* significant at 10%; ** significant at 5%; *** significant at 1%
Data are at the teacher by year level and the main outcome of interest is the proportion of students in a teachers class in a given year that is taking at least one other class at the honors level.

Table 4: *Illustration of the Variation at a Hypothetical School*

		Track A	Track B
		Alg I (regular) Eng I (regular) Natural Sciences US History	Alg I (regular) Eng I (regular) Biology World History Geometry
	Year		
Math Teacher 1	2000	X	X
Math Teacher 2	2000		X
Math Teacher 1	2005	X	X
Math Teacher 2*	2005	-	-
Math Teacher 3	2005		X

Table 5: *Variability of Teacher Effects*

	1	2	3	4	5	6	7	8	9	10
	School Effects+ lagged test scores	School Effects + student covariates	School Effects + student covariates + Peer covariates	Track-by-School Effects + student covariates +Peer covariates	Track-by-School Effects + student covariates + Peer covariates + school-by-year effects	School Effects+ lagged test scores	School Effects + student covariates	School Effects + student covariates + Peer covariates	Track-by-School Effects + student covariates +Peer covariates	Track-by-School Effects + student covariates + Peer covariates + school-by-year effects
	Algebra I					English I				
SD of raw mean teacher level residuals	0.232	0.231	0.230	0.153	0.134	0.159	0.158	0.154	0.101	0.084
F-test: Prob(all teacher level means equal)	0.000	0.000	0.000	0	0	0	0	0	0	0
SD of teacher year effects (within teachers)	0.351	0.350	0.346	0.217	0.183	0.28	0.279	0.273	0.192	0.140
Average number of years per teacher	3.10	3.10	3.10	3	2.9	3.6	3.6	3.6	3.5	3.4
SD of mean of transitory variability	0.199	0.199	0.197	0.125	0.107	0.147	0.147	0.144	0.103	0.076
Implied true SD of persistent teacher	0.119	0.117	0.119	0.087	0.080	0.061	0.058	0.055	0	0.036
Cov(t,t-1)	0.13324	0.13335	0.12356	0.01360	0.01025	-0.0059	-0.0061	-0.01057	-0.01271	-0.00124
t:test: Prob[cov(t,t-1)]>0	0	0	0	0.08	0.054	0.53	0.55	0.42	0.51	0.67
Bootstrap 95% Lower Bound: cov(t,t-1)	0.05915	0.05420	0.04656	-0.00919	-0.00445	-0.02665	-0.02511	-0.03161	-0.02318	-0.00674
Bootstrap 95% Upper Bound: cov(t,t-1)	0.20541	0.20452	0.18803	0.03638	0.02885	0.01775	0.03137	0.01723	0.00343	0.00453
Lower (sigma)	0.243	0.233	0.216	0.000	0.000	0	0	0	0	0
Implied (sigma)	0.365	0.365	0.352	0.117	0.101	0	0	0	0	0
Upper (sigma)	0.453	0.452	0.434	0.191	0.170	0.133	0.177	0.131	0.059	0.067

Student covariates include 8th grade and 7th grade math scores (and their third order polynomials) , 8th grade and 7th grade math scores (and their third order polynomials), parental education, ethnicity, gender, LEP status, number of honors courses taken, and whether the students has been designated gifted in English and/or math. Peer covariates are the classroom level and school means of all the student level covariates.

Table 6: *Out of sample predictions*

	Algebra Teacher Standardized Score (2008-2010)					Algebra Teacher Standardized Score (2008-2010)				
	Using both variation within track-school-year cells and variation within track school cells across years		Using only variation within track-school-year cells	Using variation within track school cells across years		Using both variation within track-school-year cells and variation within track school cells across years		Using only variation within track-school-year cells	Using variation within track school cells across years	
	1	2	3	4	5	6	7	8	9	10
Standardized Teacher Effect (2005-7)	**0.03521**	**0.02381**	**0.02426**			**0.01027**	**0.01246**	**0.01385**		
	[0.00632]* **	**[0.00677]*** **	**[0.00717]*** **			**[0.00392]*** **	**[0.00330]*** **	**[0.00353]*** **		
Mean Standardized Teacher Effect (2005-7)				**0.03861**	**0.0345**				0.00431	0.00034
				[0.00886]* **	**[0.0190]* **				[0.00399]	[0.00725]
School Effects	Y	-	-	-	-	Y	-	-	-	-
Track-School-Effects	N	Y	-	Y	Y	N	Y	-	Y	Y
School-Year-Effects	N	Y	-	N	N	N	Y	-	N	N
School-Track-Year Effects	N	N	Y	N	N	N	N	Y	N	N
Student and peer covariates	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	48195	45874	47508	67036	64447	104648	99854	102864	129024	123624

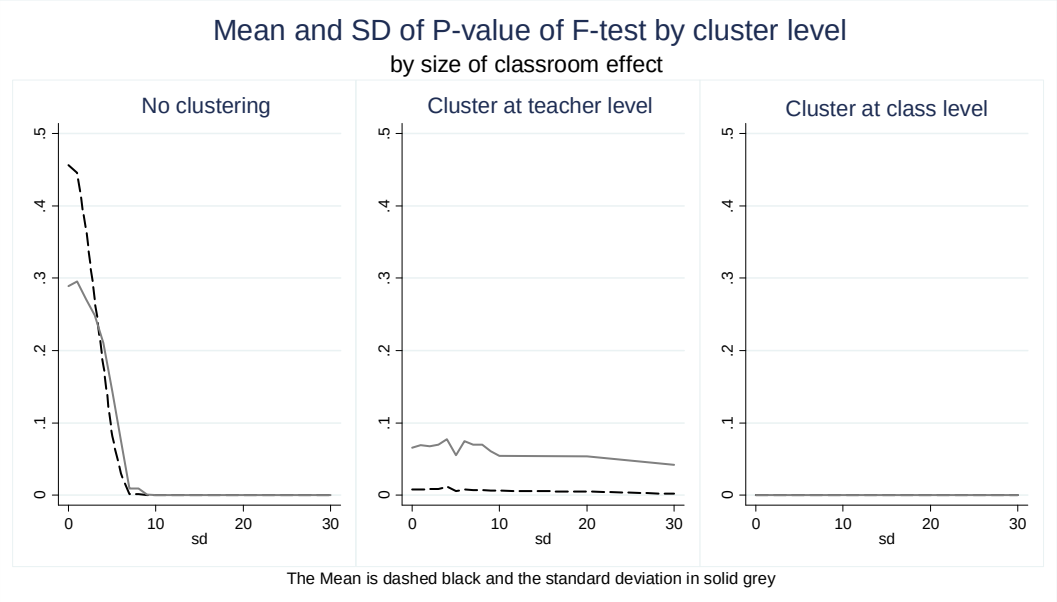
Robust standard errors in brackets

* significant at 10%; ** significant at 5%; *** significant at 1%

Student covariates include 8th grade and 7th grade math scores (and their third order polynomials) , 8th grade and 7th grade math scores (and their third order polynomials), parental education, ethnicity, gender, LEP status, number of honors courses taken, and whether the students has been designated gifted in English and/or math. Peer covariates are the classroom level and school means of all the student level covariates.

Appendix

Figure A1: *F-tests with clustering (no classroom effects and no teacher effects)*



The size of the transitory teacher-by-year level shock is indicated on the X-axis.

Figure A2: *Density of Test Score Outcomes*

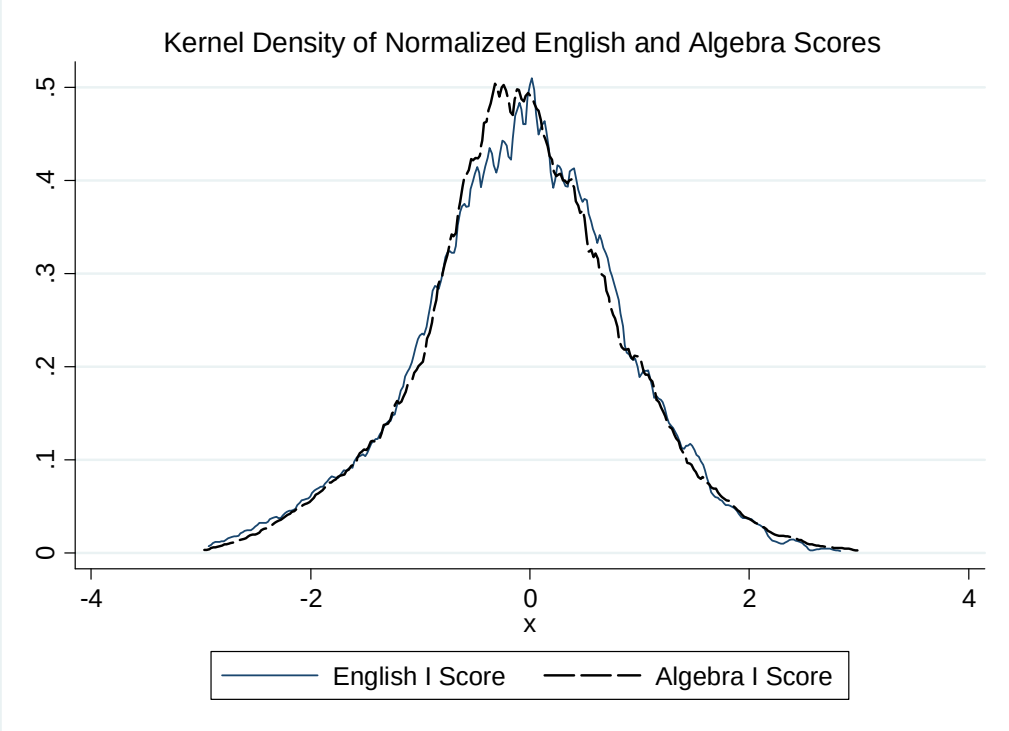


Table A1: *Distribution of Number of Teachers in Each School-Track Year Cell*

	Percent	
Number of Teachers in Track-Year-School Cell	English	Algebra
1	63.37	51.07
2	18.89	26.53
3	9.12	11
4	5.6	6.38
5	3.03	3.25
6	0	1.77

Note: This is after removing singleton tracks.

Table A2: *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**

Note: The estimated teacher effects are based on a value-added model that does not include school-track effects, but does include school fixed effects, year fixed effects, and all student and peer covariates.

Appendix Note 1: *Matching Teachers to Students*

The teacher ID in the testing file corresponds to the teacher who administered the exam, who is not always the teacher that taught the class (although in many cases it will be). To obtain high quality student-teacher links, I link classrooms in the End of Course (EOC) testing data with classrooms in the Student Activity Report (SAR) files (in which teacher links are correct). The NCERDC data contains The End of Course (EOC) files with test score level observations for a certain subject in a certain year. Each observation contains various student characteristics, including, ethnicity, gender, and grade level. It also contains the class period, course type, subject code, test date, school code, and a teacher ID code. Following Mansfield (2011) I group students into classrooms based in the unique combination of class period, course type, subject code, test date, school code, and the teacher ID code. I then compute classroom level totals for the student characteristics (class size, grade level totals, and race-by-gender cell totals). The Student Activity Report (SAR) files contain classroom level observations for each year. Each observation contains a teacher ID code (the actual teacher), school code, subject code, academic level, and section number. It also contains the class size, the number of students in each grade level in the classroom, and the number of students in each race-gender cell.

To match students to the teacher who taught them, unique classrooms of students in the EOC data are matched to the appropriate classroom in the SAR data. To ensure the highest quality matches, I use the following algorithm:

(1) Students in schools with only one Algebra I or English I teacher are automatically linked to the teacher ID from the SAR files. These are perfectly matched. Matched classes are set aside.
(2) Classes that match exactly on all classroom characteristics and the teacher ID are deemed matches. These are deemed perfectly matched. Matched classes are set aside.
(3) Compute a score for each potential match (the sum of the squared difference between each observed classroom characteristics for classrooms in the same school in the same year in the same subject, and infinity otherwise) in the SAR file and the EOC data. Find the best match in the SAR file for each EOC classroom. If the best match also matches in the teacher ID, a match is made. These are deemed imperfectly matched. Matched classes are set aside.
(4) Find the best match (based on the score) in the SAR file for each EOC classroom. If the SAR classroom is also the best match in the EOC classroom for the SAR class, a match is made. These are deemed imperfectly matched. Matched classes are set aside.
(5) Repeat step 4 until no more high quality matches can be made.

This procedure leads to a matching of approximately 60 percent of classrooms. All results are similar when using cases when the matching is exact so that error due to the fuzzy matching algorithm does not generate any of the empirical findings.

Appendix Note 2

To provide evidence of bias due to sorting on observables, I determine if teacher value-added is correlated with observable student characteristics. I estimate teacher value-added using data from 2005 through 2007[16], 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.[17] 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 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 A3. 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 used 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 A3. For Algebra the point estimates are very small, and none of the coefficients is statistically significant at the 5 percent level. In contrast, while the estimate is not statistically significant for English teachers on predicted English scores, the point estimates is sufficiently large (0.074) that selection to English teachers within tracks *might* be a concern.

Table A3: *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]	[0.0688]	

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

[16] I compute Empirical Bays estimates following Kane and Stagier 2008. See appendix note 1.
[17] 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.

Appendix Note 3 *Evidence of student sorting to teachers*

The evidence thus far indicates that is sorting of student and teachers to tracks. However, one may wonder about student sorting to teachers directly. To asses this, following Aaronson, Barrow, and Sander (2007), 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 in a given year) observed in the data and compare that to the mean within-teacher student test score dispersion for other counterfactual assignments. Table A4 displays the actual within teacher-year test score dispersion, what one would observe with full student sorting to teachers within schools, and random student assignment to teachers within schools.[18] 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. However, in a Monte Carlo simulation of mean test score dispersion under random assignment within schools, none of the 500 replications yielded dispersion levels as low as that observed ─ suggesting that there is some systematic sorting of student to teachers based on incoming achievement.

Table A4: *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.

[18] For the interested reader, I also present the test score dispersion with full student sorting across all classroom and schools and random student assignment across all classrooms and schools.

Appendix Note 3: *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. 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 construct an estimate of the estimation error for each teacher, I first estimate the variance of the year-to year variation within teachers over time (we will call this $\sigma_{c_{jy}}^2$). This is an estimate of the variability of transitory teacher by year shocks and sampling variability. Then to account for the fact that teachers observed in more years will have more reliable estimates I estimate $\sigma_{u_j}^2$ (the estimation variance of the raw value added estimate) with $\sigma_{c_{jy}}^2 / C_j$ where C_j is the number of years observed for teacher j. I take as my estimate of σ_θ^2 the implied standard deviation of teacher effect from Table 7. Using the covariance of mean residuals across years for the same teacher as an estimate of σ_θ^2 yields almost identical estimated for Algebra teachers. However, because the estimated covariance is negative for English teachers, one can only use the adjusted implied variance. To obtain an EB estimate for each teacher, I multiply the raw teacher effect $\hat{\theta}_j$ by an estimate of its reliability. Specifically, I compute $\hat{\theta}_j^{EB} = \overline{\theta}_j \cdot \hat{\sigma}_\theta^2 / (\hat{\sigma}_\theta^2 + \sigma_{u_j}^2)$ .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 $\hat{\theta}_j$ is for θ_j.