October, 2014

Teacher Quality at the High-School Level: The Importance of Accounting for Tracks

C. Kirabo Jackson, *Northwestern University*

Teacher Quality at the High-School Level: The Importance of Accounting for Tracks

C. KIRABO JACKSON[1]
MAY 17, 2013
Northwestern University, IPR, and NBER

Unlike in elementary school, high-school teacher effects may be confounded with both selection to tracks and track-level treatments. I document confounding track effects, and show that traditional tests for the existence of teacher effects are biased. After accounting for biases, high-school algebra and English teachers have much smaller test-score effects than found in previous studies. Moreover, unlike in elementary school, value-added estimates are weak predictors of teachers' future performance. Results indicate that either (a) teachers are less influential in high school than in elementary school, or (b) test-score effects are a poor measure of teacher quality at the high-school level. (JEL I21, J00).

There is a current push to use test-score based measures of teacher quality in policy decisions. Several districts including Washington D.C., New York, and Los Angeles publicly release estimates of a teacher's average effect on her students' test scores (value added) and many districts have begun using them in hiring, firing, and retention decisions. The increasing reliance on test-score based measures of teacher quality is predicated primarily on the notion that one can identify teachers who will improve test scores in the future based on how well they increased test scores in the past — i.e. that estimated teacher value added is persistent over time. This paper demonstrates that due to previously unaccounted-for biases, existing studies on teacher quality, particularly those at the high-school level, may overstate the magnitude of *persistent* teacher effects on test scores, and may drastically overstate their usefulness for policy.

The emphasis on test-score measures of teacher quality is based largely on studies of elementary-school teachers showing that a one standard deviation increase in teacher quality leads to between 0.1 and 0.2 of a standard deviation (σ) increase in math and reading scores (e.g. Rivkin et. al. 2005; Rockoff, 2004; Kane & Staiger, 2008). Using similar methodologies as those used for elementary-school teachers, the few studies on high-school teachers find similar and somewhat larger effects. Using value-added models, Aaronson, Barrow, and Sander (2007), Koedel (2008), Goldhaber et. al. (2011) and Slater, Davies, and Burgess (2011) find that a standard deviation

[1] Early versions of this paper were previously circulated as "Do High-School Teachers Really Matter?" I would like to thank David Figlio, Jon Guryan, and seminar participants at Georgetown University, the University of Kentucky, the Cleveland Fed, Uppsala University, and Leibnitz University for helpful comments. All errors are my own.

improvement in high school teacher quality raises student scores by between 0.10σ and 0.4σ.[2]

In many nations elementary-school students are exposed to a single teacher in a given year and follow the same course of study, while secondary-school students are exposed to several teachers, take different courses, and placed into different tracks. This difference may render methodologies appropriate for measuring elementary-school teacher quality inappropriate for other grades for two reasons. First, there may be unobserved differences between students taking the same course with similar incoming test scores but who are in in different tracks (group of courses). For example, taking a greater number of honors courses (among all the courses taken) may be associated with greater motivation even among students with the same past performance and who take the same math or English course. With such selection to tracks, if different teachers of the same course teach students from different tracks, there may be bias due to selection to tracks that will not be accounted for with controls for past performance or the covariates used in standard value-added models. Second, distinct from selection bias, there is a possible omitted variables bias. Specifically, in most high-school settings, *even with random assignment of students to teachers,* if different teachers teach in different tracks (group of courses), 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 taking Algebra I with Mr. Smith take physics with Mr. Black who affects algebra scores directly), the content of other courses (e.g. students taking Algebra I with Mr. Smith also take physics which affects algebra scores directly), or track-specific treatments (e.g. college-bound students take Algebra I with Mr. Smith and are also part of a program that teaches study skills that affect algebra scores directly). Existing studies have not addressed the omitted variables bias associated with track-specific treatments or the specific forms of selection that may occur at the high-school level.[3] As such, it remains unclear to what extent high-school teachers affect test scores, and whether test-score measures of teacher quality are a useful policy tool at the high-school level.

I aim to (a) demonstrate that the statistical methods used to identify teacher test-score effects in elementary school may be inappropriate for identifying teacher test-score effects at the middle- or high-school level, (b) use data on high-school algebra and English teachers in North

[2] Clotfelter, Ladd, and Vigdor (2010) find that teacher characteristics are associated with test score differences.
[3] Two recent working papers, Harris and Anderson (2013), and Protik, et al. (2013) find that failing to account for course tracks leads to sizable biases in the individual value-added estimates for middle- and high-school teachers.

Carolina to present evidence on the existence of track-specific effects, (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, and (d) document the extent to which historically-estimated value added predicts teacher effectiveness at improving test scores in a high-school setting, and compare these results to those found for elementary-school teachers.

To address the biases associated with tracks, 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 to schools, tracks, and courses. 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. The remaining concern is that comparisons among students within the same track may still 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.

This paper also makes two methodological contributions to the broader teacher-quality literature. The first contribution is to show that in most datasets where individual teachers are not observed in several classrooms (i.e. more than 20), the commonly used F-test will yield very small p-values even when there are no real teacher effects. Motivated by Kane and Staiger (2008), I present an unbiased test for the existence of teacher effects that does not require a large number of classroom observations for each teacher, but only requires observing a large number of teachers in multiple classrooms (a condition satisfied in most datasets). The second methodological contribution is to present an approach to estimating the variance of the variance of teacher quality effects. Under certain conditions, the covariance across classroom effects for the same teacher is a consistent estimate of the variance of true teacher quality. Because a covariance is a sample statistic, one can obtain confidence intervals for the variance of true teacher quality. While existing

studies present estimates of the variance of teacher quality effects, they do not present confidence intervals for these estimates so that this approach may be a useful contribution.

I document that traditional value-added models yield a strong positive association between teacher experience and student outcomes (similar to existing studies). However, in models that compare teachers within the same track, there is a substantively weaker association between test scores and teacher experience in both subjects. Consistent with this, after accounting for track effects, the plausible range of estimated standard deviations of algebra and English teacher effects are smaller than estimates from other studies. Specifically, the 95% confidence intervals for the standard deviation of teacher effects is (0.0514σ ; 0.081σ) and (0 ; 0.0312σ) for algebra and English, respectively. Looking to usefulness for policy, I find that the ability to predict future performance based on historical performance at improving test scores is much smaller than at the elementary level. Jackson and Bruegmann (2009) find that a good elementary-school teacher (85 percentile) based on previous performance raises current scores by 0.17σ and 0.08σ more than an average teacher in math and English, respectively. In stark contrast, a good 9th grade teacher (85 percentile) based on previous performance raises current scores by only about 0.027σ and 0.018σ standard deviations more than an average teacher in algebra and English, respectively. The results suggest that (a) the properties of high- and elementary-school teacher value added are meaningfully different, (b) teachers may be less influential in high school than in elementary school, and (c) using test scores to identify effective teachers may have limited practical benefits at the high-school level. Note that that this study focuses on policy-relevant persistent teacher effects that can be reasonably attributed to teachers (rather than transitory variability that may reflect random shocks). Also, this study speaks only to teacher effects on student skills captured by standardized tests.[4]

The remainder of the paper is as follows: Section II describes the data, Section III details the econometric 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 include

[4] See Jackson (2012) for an analysis of teacher effects on non-test score outcomes.

demographic data, middle-school achievement data, end of course data (high-school), 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 ninth graders take at least one of these courses so that the resulting sample is representative of ninth 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.

These data cover 348,547 ninth grade students in 619 secondary schools in classes with 4296 English I teachers, and 3527 Algebra I teachers. While roughly half of the students are male, about 58 percent are white, 29 percent are black, 7.5 percent are Hispanic, 2 percent are Asian, and the remaining one percent is Native American, mixed race, or other. About 7.5 percent of students have the highest level of education of the student's two parents below high-school, 40 percent with a high school degree, 15 percent with a junior college or trade school degree, 20 percent with a four year college degree or greater, and 6.4 percent with an advanced degree (the remaining 10 percent are missing parent education data). Test score variables have been standardized to be mean zero with unit variance for each cohort and test. All tests have single peaked bell-shaped distributions, and there is no evidence of any ceiling or floor effects.[6]

Measuring Tracks:

It is well documented that most high schools (even those with no explicit tracking policy) practice de-facto tracking by placing students of differing levels of perceived ability (e.g. college bound, remedial, honors) into distinct groups of courses (Sadker and Zittleman 2006, Lucas and Berends 2002). In North Carolina, while there are hundreds of courses that students can take (including special topics, physical education, and extracurricular activities) there are 10 core

[5] The data link students to the teacher who administered the test. In most cases this is the student's 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 B for details on the matching procedure. The matched sample corresponds to 57 percent of all Algebra student observations and 69 percent of all English student observations. The demographic characteristics of students in matched classroom are very similar to those in unmatched classrooms. However, unmatched classrooms appear to have somewhat higher incoming student test scores than the full sample. All results are similar when using cases when the matching is exact.

[6] Incoming 8th grade test scores in the final 9th grade sample are approximately 8 percent of a standard deviation higher than that of the average in 8th grade. This is because the sample of 9th grade students is less likely to have repeated a grade and to have dropped out of the schooling system.

academic courses (listed in Table 2) that make up over half of all courses taken. Because Algebra I and English I are taught at multiple levels (advanced, regular, and basic), students in the "high-ability" track take these courses at the advanced level, while those in the "lower ability" tracks will take these courses at the basic level. Because the advanced class might emphasize different material from the regular class, *even with random assignment of students to classes*, basic comparisons of outcomes across 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 the 10 largest 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.[7] Students who take the same 10 most commonly taken courses at different schools are in different school-tracks. Students at the same school who take either a different number of the 10 courses, or at least one different course (among the 10 courses) are in different school-tracks. In addition, students at the same school who take the same courses (among the 10 courses) but took either Algebra I, or English I, at different levels are in different school-tracks. Because many students take courses in groups (Frank, et al. 2008), only 3.7 percent of students are in singleton tracks, most are in school-tracks with more than 50 students, and the average school-track contains 117 students. There are 14,793 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.

Comparing the standard deviations within schools to those within school-tracks provides evidence of sorting into tracks. If there were no sorting to tracks within schools, these two estimated standard deviations would be similar. For all 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 0.878 while it is only 0.600 within school-tracks (a 33% reduction in the variance). Consistent with these tracks capturing important differences between

[7] Note that this measure of tracks is stable over time and is likely not endogenous to teacher quality. This would be the case if students often switched tracks during the year in response to teacher quality. The data show that such switching half way through the year is relatively rare. First about two-thirds of students are at schools that teach Algebra I and English I over one semester meeting twice as often rather than over two semesters. At the schools that do teach Algebra and English over the course of a year, 2.8 and 4.6 percent of students switch the level of Algebra I and English I taken between the fall and the spring, respectively. Accordingly, track switching appears to be a relative rare occurrence and should not bias the results in any appreciable way.

college-bound, remedial, and average students, the standard deviation of the number of honors courses taken is 1.163 within schools while it is only 0.575 within school-tracks (a 50% 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 tracks and groups students by ability.

III Econometric Framework:

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

[1] $$Y_{ijcgy} = A_{iy\text{-}1}\delta + X_{iy}\beta + I_{ij}\cdot\theta_j + \pi(P|g) + \mu_{jc} + \varepsilon_{ijcgy.}$$

Here, $A_{iy\text{-}1}$ is incoming achievement level of student i, X_{iy} is a matrix of student-level covariates, I_{ij} is an indicator variable equal to 1 if student i has teacher j 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_{jc} is a random within-teacher classroom-level error such that $E[\mu_{jc}|J]=0$, and ε_{ijcgy} is a random mean zero error term. When track-level treatments are unobserved and one does not account for classroom errors, the OLS estimate of teacher effect $\hat{\theta}_j$ is given by [2].

[2] $$[\hat{\theta}_j \mid J, X, A] = \theta_j + \left(\sum_{g=1}^{G} w_{jg}\cdot\left(\sum_{i\in g}\frac{e_{ijcgy}}{N_g}\right)\right) + \left(\sum_{g=1}^{G} w_{jg}\cdot\pi(P\mid g)\right) + \left(\sum_{c\in j}\frac{\mu_{jc}}{T_j}\right) + \left(\frac{1}{N_j}\sum_{i\in j}\left(e_{ijcgy} - \sum_{i\in g}\frac{e_{ijcgy}}{N_g}\right)\right).$$

In [2], w_{jg} is the proportion of students who are in class with teacher j that are also in track g, N_g is the number of students in track g, N_j is the number of students with teacher j, and T_j is the number of classrooms observed for teacher j. I discuss the components of this equation below.

Potential Bias in Estimated Teacher Effects:

Equation [2] highlights that there are four distinct potential sources of bias in estimated teacher effects. The first source is $A \equiv \left(\sum_{g=1}^{G} w_{jg}\cdot\left(\sum_{i\in g}\frac{e_{ijcgy}}{N_g}\right)\right)$, which is a weighted average of the student mean error term in each track (where the weight on track g for teacher j is the proportion of students with teacher j that are in track g). This term 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 term $B \equiv \left(\frac{1}{N_j} \sum_{i \in j} \left(e_{ijcgy} - \sum_{i \in g} \frac{e_{ijcgy}}{N_g} \right) \right)$ which is the average track de-meaned student-level error for teacher *j,* and represents sampling variability. If each teacher is observed with more than 30 students (as is often the case), this term will be small and will result in minimal bias. A third potential source of bias is the term $C \equiv \left(\sum_{c \in j} \frac{\mu_{jc}}{T_j} \right)$, which is the mean of the transitory classroom-level disturbances for a given teacher (assuming equally-sized classrooms). This source of bias, discussed in Kane and Staiger (2008), is due to transitory shocks that may be correlated with individual teachers *in finite samples*. For example, a teacher who teaches a class that had a flu outbreak will have lower performance than expected and, *based on only one class*, will appear to be less effective than her true ability. While this term should be near zero if teachers are observed in many classrooms, this bias term will be non-trivial and non-zero in short panels (small number of classes per teacher). In elementary school, most teachers teach one class per year so that one would need panels with close to thirty years of data per teacher for this term to be negligible. In high school, most teachers teach between one and three classes per year so that one would need at least 10 years of data for each teacher for this term to be negligible. Most existing datasets do not provide such long panels.

The potential source of bias that has not been addressed by the existing literature is the term $D \equiv \left(\sum_{g=1}^{G} w_{jg} \cdot \pi(P \mid g) \right)$. This is a weighted average of the unobserved treatments (e.g. extra tutoring/mentoring or time management classes) in each track (where the weight on track *g* for teacher *j* is the proportion of students with teacher *j* that are in track *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 other honors classes teach students study skills that directly affect their algebra outcomes, one might erroneously attribute the benefits from the additional study-skills training in the honors track to Mr. Jones. This is a form of omitted variables bias that will exist *even if students are randomly assigned to teachers* so long as individual teachers are correlated with certain track-specific treatments. Section V shows that this is the case in North Carolina high schools.

The main objective of this study is to estimate the variability of persistent teacher quality. As explained above, the variance of the estimated teacher effects $\hat{\theta}$ from [2] may overstate the variance of true persistent teacher quality because (a) this confounds teacher effects with track-level selection and treatments, and (b) teacher effects are estimated with error due to sampling variation and transitory classroom-level disturbances. I propose strategies to address this in Section IV.

IV Empirical Strategy

IV.1 *Estimating the Variance of Persistent Teacher Quality*

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, 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 [3] below.

[3] $$Y_{ijcgy} = A_{iy\text{-}1}\delta + X_{iy}\beta + I_{jy}\cdot\theta_j + I_{gi}\,\theta_g + \mu_{jc} + \varepsilon_{ijcgy.}$$

By conditioning on school-tracks, one can obtain consistent estimates of the teacher effects θ_j as long as there is no selection to teachers *within* a school-track.[8]

Because the main models include school-track effects, teacher effects are identified by comparing outcomes of students at the same school in the same track but who have different teachers. There are two sources of identifying variation:(1) comparisons of teachers at the same school teaching students in the same track *at different points in time*; and (2) comparisons of teachers at the same school teaching students in the same track *at the same time*. To illustrate these sources of variation, consider the simple case illustrated in Table 3. There are two tracks A and B in a single school. There are two math teachers at the school at all times.

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 effect 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 track B with

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

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. As long as teacher 1 does not become more effective by diverting resources away from other teachers toward her own students, this difference will reflect school-wide changes in outcomes that will have the same effect on all teachers. In a regression setting this is accomplished with the inclusion of school-by-year fixed effects. 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 Section V and find no evidence of selection to cohorts.

The second source of variation comes from having multiple teachers teaching the same course in the same school-track at the same time. In the example above, because both teachers 1 and 2 teach students in track B in 2000, one can estimate the effect of teacher 1 relative to 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 if students do not select to teachers *within* tracks. I test for this in Section V and find no evidence of selection to teachers 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, I show that the results are robust to using each distinct source of variation separately.

Removing bias due to within-teacher classroom-level shocks and sampling variability

While the variance of the estimated teacher effects $\hat{\theta}$ from [3] will no longer 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 separate approaches to address this issue.

Results using both approaches are very similar.

If any teacher-level variation in test scores not explained by classroom shocks or sampling variability is due to teacher quality, and classroom shocks or sampling variability are unrelated to true teacher quality, one can compute the variance of the raw teacher effects and subtract the estimated variance of the transitory classroom-level shocks (Kane and Staiger 2008).[9] I estimate the combined variance of the classroom shocks and sampling variation with the variance of mean classroom-level residuals *within teachers* (i.e. the within-teacher variability in average test scores over time). I then divide this by the number of observed classrooms for each teacher to obtain an estimate of the variability in the raw teacher effects that can be attributed to transitory variability. Because this approach attributes any variability not explicitly accounted for by classroom effects or sampling variability to teachers, it may overstate the true variance of persistent teacher quality. This motivates the second approach outlined below.

In the second approach to estimating the variance of teacher quality I follow Kane and Staiger (2008). I estimate [3] without teacher indicator variables, and take the covariance across mean classroom-level residuals for the same teacher as my estimate of the variance of the *persistent* component of teacher quality (i.e. that observed across classrooms). This is achieved in two steps:

<u>*Step 1:*</u> Estimate equation [4] below.

[4] $$Y_{igcjy} = A_{iy\text{-}1}\delta + X_i\beta + X^*_{jc}\pi + I_{gi}\,\theta_g + \theta_{sy} + \varepsilon^*_{ijcgy}$$

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 student covariates (parental education, ethnicity, and gender). X_{jc}* is the mean characteristics of other students in class c with teacher j. 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 classroom effect so that $\varepsilon^*_{ijcgy} = \theta_j + \mu_{jc} + \varepsilon_{ijcgy}$.

<u>*Step 2:*</u> Link every classroom-level mean residual $\bar{e}^*_{jc}$, pair it with a random different

[9] Formally, if all the terms in [2] are uncorrelated, then after removing variation due to tracks (i.e terms A and D) we have that $Var(\hat{\theta}) = \sigma^2_\theta + \sigma^2_B + \sigma^2_C$ so that $\sigma^2_\theta = Var(\hat{\theta}) - \sigma^2_B - \sigma^2_C$.

classroom-level mean residual for the same teacher $\bar{e}^{*}_{jc'}$ and compute the covariance of these mean residuals (i.e. compute $cov(\bar{e}^{*}_{jc}, \bar{e}^{*}_{jc'})$), where each classroom is weighted by the number of students in the classroom. To ensure that the estimate is not driven by any particular random pairing of classrooms within teachers, I replicate this procedure 200 times and take the mean of the estimated covariances as the parameter estimate. If the classroom errors μ_{jc} are uncorrelated with each other and are uncorrelated with teacher quality θ_j, the covariance of mean residuals *within teachers* but *across classrooms* is a consistent measure of the true variance of persistent teacher quality. That is, $cov(\bar{e}^{*}_{jc}, \bar{e}^{*}_{jc'}) = \text{cov}(\theta_j, \theta_j) = \text{var}(\theta_j) \longrightarrow \sigma^2_{\theta_j}$.[10]

Because these procedures remove a track-specific mean in the first stage, they may remove some signal about teacher quality. This problem is most severe when there are a small number of teachers in a track (in the extreme, there is no remaining signal for the sole teacher in a track). Appendix B shows that removal of a track specific mean may lead one to understate the variance of teacher quality by a factor of (1-1/R) where *R* is the number of teachers in a track. As such, if all tracks have only 2 teachers the raw estimated variance after removing track effects will be roughly 0.5 of the actual variance. Similarly, if all tracks have 10 teachers the estimated variance after removing track effects will be roughly 0.9 of the actual variance. Table A2 shows the proportion of the students that have a given number of teachers in the track and the associated degree of freedom adjustment (i.e. 1/(1-1/R)). Taking the average over all the data, the implied degree of freedom adjustment is 1.239 for algebra and 1.244 for English. Accordingly, to avoid erroneously attributing a mechanical reduction in variance to track effects, in models that remove school-track effects I inflate the estimated variance by 1.239 for algebra and 1.244 for English.

IV.2 *Statistical Inference for Teacher Effects*

Testing for the existence of Teacher Effects

While estimating the variance of teacher effects in a particular sample is important, being able to determine whether such estimates are statistically significant (or occurred by random chance) is equally important. This section illustrates that commonly used tests for the existence of teacher effects suffer from a severe finite sample bias in most existing datasets that do not contain

[10] Note: $\text{cov}(\theta_j, \mu_{jc'}) = \text{cov}(\theta_j, \bar{e}_{jgyc'}) = \text{cov}(\theta_j, \mu_{jc}) = \text{cov}(\mu_{jc}, \mu_{jc'}) = \text{cov}(\mu_{jc}, \bar{e}_{jgyc'}) = \text{cov}(\bar{e}_{jgyc}, \theta_j) = \text{cov}(\bar{e}_{jgyc}, \mu_{jc'}) = \text{cov}(\bar{e}_{jgyc}, \bar{e}_{jgyc'}) = 0$

teachers in several classrooms. I then present an alternate test that provides unbiased statistical inference in datasets with a large number (>30) of teachers observed in more than one classroom—a condition satisfied in most existing datasets.

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 mean outcomes are the same across all teachers. Formally, researchers test $H_0 : \theta_1 = \theta_2 = \theta_3 = ... = \theta_J$. The *F*-statistic is the ratio between the across- and within-teacher variance in test scores.[11] If the variance across teachers is large relative to the within-teacher variability in outcomes, this statistic will be large and it would imply that there is something occurring at the teacher level that explains variability in outcomes (i.e. that teacher effects are not all equal).

To illustrate how the existence of transitory classroom shocks affects the F-test, let us abstract away from selection bias, track treatments, and sampling variability. In this case [2] simplifies to $\hat{\theta}_j = \theta_j + \sum_{j=1}^{T_j} (\mu_{jc} / T_j)$. Now the test of equality of teacher-level mean outcomes is really a test of equality of the teacher effects *plus the means of the transitory classroom shocks for each teacher*. This test is asymptotically equivalent to testing the equality of teacher effects only when *each* teacher is observed in several classrooms so that $\sum \left(\mu_{jy} / T_j \right) \longrightarrow 0 \ \forall j$. However, if teachers are observed in a handful of classrooms, one cannot ignore the influence of transitory classroom shocks. In the extreme (but not uncommon) case where each teacher is observed in only one classroom, it is impossible to distinguish between a good (bad) teacher and teacher who happened to have a good (bad) classroom shock. As such, with unaccounted-for classroom shocks, the *F*-test will tend to over-reject the null hypothesis even when it is true.

To get a sense of the behavior of the commonly used *F*-test I implemented the following procedure. Using the actual data I (1) drew a random error for each actual classroom from a mean zero normal distribution, (2) drew a random student-level error for each actual student from a standard normal distribution, (3) created a simulated outcome for each student that is the sum of these two errors, and (4) tested for the existence of teacher effects using the F-test (note there are no teacher effects in the simulated outcome). I repeated this procedure 500 times each for

[11] The F-statistic is $F = \frac{1}{J-1} (\hat{\theta} - \bar{\hat{\theta}})' (\hat{V}_J)^{-1} (\hat{\theta} - \bar{\hat{\theta}}) \equiv \frac{\text{[across-teacher variance]}}{\text{[within-teacher variance]}}$, where $\hat{\theta}$ is a J×1 vector of estimated fixed effects, $\bar{\hat{\theta}}$ is the mean of the $\hat{\theta}_j$s, and $\hat{V}_J$ is the J×J variance-covariance matrix for the teacher effects.

classroom shocks of different variability. Figure 1 shows the distribution of the *p*-values associated with the *F*-tests for varying variability of the classroom-level shocks. Because there are no teacher effects, if the test is unbiased, the *p*-values should follow a uniform distribution centered on 0.5.

With no classroom errors (sd=0.00) the distribution of *p*-values is roughly uniform—indicating that the test is unbiased in the absence of classroom-level shocks. However, with quite modest classroom errors (sd=0.05) the likelihood of a *p*-value smaller than 0.1 is about 70 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 ubiquitously used *F*-test virtually uninformative. The small number of clusters per teacher is a finite sample problem that is not solved by clustering the standard errors, degree of freedom adjustments, or aggregating the data (Bertrand, Duflo, and Mullainathan 2004; MacKinnon and White 1985).[12] As such, existing studies that rely on the *F*-test likely overstate the degree to which we can be sure that teacher effects exist.

An unbiased test for the existence of teacher effects:

To address the problems above, I propose a new test based on the idea that if teacher effects exist, the residuals for a teacher in one classroom should be correlated with her residuals from another classroom. One can test this by running a regression of a teacher's mean residuals in classroom *jc* on her mean residuals in classroom *jc'* and then implementing a simple *t*-test on the coefficient on classroom *jc'* residuals (for the same teacher). I do this using a single draw of a random pairing of each classroom with another classroom for the same teacher. As long as classroom disturbances are uncorrelated, this test 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 classroom *jc* and *jc'*) for all teachers on average rather than one parameter for each teacher. Because the *t*-test is a finite sample test, this new test will yield valid statistical inference on samples with *any* number of teachers as long as at least two teachers are observed in more than one classroom. This is satisfied in most existing datasets.

To illustrate the unbiasedness of this proposed test, I show the covariance test's

[12] Because the *F*-test requires that the each estimated coefficient be normally distributed, and asymptotic normality cannot be invoked with a small number of independent observations, the F-test will not be reliable in short panels.

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 500 replications for classroom disturbances of different variability. If the test is unbiased, the *p*-values should follow a uniform distribution centered around 0.5. As one can see, irrespective of the size of the classroom errors, the *p*-values follow a uniform distribution, so that this test is robust to idiosyncratic classroom disturbances and will be an unbiased test for the existence of teacher quality effects. I will use this test.

Deriving confidence bounds for the variance of teacher quality effects

Because existing studies do not present confidence intervals for estimates of the variance of teacher effects, one cannot meaningfully compare estimates across studies or know what range of effect sizes is consistent with the data. Given the increasing use of estimates from studies to inform policy, it is important to be able to report the degree of uncertainty around these estimates. Addressing this need in the literature, I propose a way to estimate confidence bounds around estimated variances of teacher effects. Because the variance of the teacher effects can be estimated by a sample covariance, one can compute confidence intervals for the variance of the persistent teacher effects as long as one knows the properties of the sampling distribution. I use the empirical distribution of 500 randomly drawn and computed "placebo" covariances (i.e. sample covariance across randomly drawn classrooms for *different* teachers) to form an estimate of the standard deviation of the sampling variability of the covariance across classrooms for the *same* teacher. I use this "bootstrapped" standard deviation of the covariance to form normal-distribution-based confidence intervals of the covariance (i.e. the true variance of teacher effects).

V Results

V.1 Evidence of sorting of teachers into tracks

The existing literature on teacher quality has emphasized the biases associated with student sorting to teacher or tracks. However, the literature has not addressed the potential biases associated with *teachers* sorting into tracks. As illustrated in Section III, even with no student sorting into tracks (and no differences in student ability across teachers), if *teachers* sort into tracks there could be substantial bias if certain tracks provide unobserved treatments that influence classroom outcomes directly. Because this type of sorting has not been previously documented, it is instructive to present evidence of teacher sorting into tracks before presenting the main results.

To test for *teacher* sorting to tracks, I computed the proportion of students in each teacher's classes in each year who took at least one 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. 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. To assess how 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 4) 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/fewer honors classes than the average student than other teachers. This relationship holds across both subjects, and holds both across and within schools.

This relationship is problematic insofar as there is selection to honors classes, or honors classes provide skills that affect Algebra I or English I scores directly. In either scenario, taking any other honors class will predict the outcomes in a traditional value-added model. To test for this, I estimate models that predict Algebra I and English I scores as a function of school effects, year effects, 8th and 7th grade scores in both English and math, and an indicator denoting whether a student took any honors course other than the course in question. In such models, the coefficient on taking any other honors course is 0.12 in Algebra and 0.14 in English (both significant at the 1 percent level). While this honors student effect may be due to selection or track-specific treatments, the effects clearly indicate the importance of explicitly controlling for tracks**.**

V.2 Evidence of Bias based on Teacher Experience Effects

It is instructive to foreshadow the main results by illustrating how conditioning on school-tracks affects the relationship between teacher experience and test scores. To do this, I regress English I and Algebra I scores on 8th and 7th grade math and reading scores and indicator variables for each year of teacher experience. I estimate models with school fixed effects, and then with track-by-school fixed effects. I plot the coefficients on the years of experience indicator variables in Figure 3. Models with school fixed effects indicate that students have higher Algebra I scores

when they have teachers with more years of experience (top). Within schools, students with a teacher with 10 years of experience score 0.1σ higher than those with a rookie teacher. The *p*-value associated with the null hypothesis that all the experience indicators are equal is 0.0001. However, models with track-by-school effects are about 30 percent smaller such that students with a teacher with 10 years of experience score only 0.07σ higher than a rookie teacher. The difference is even more pronounced for English teachers. Models with school fixed effects indicate that students have higher English I scores when they have teachers with more years of experience (bottom). Within schools, students with a teacher with 10 years of experience score 0.03σ higher than those with a rookie teacher. The *p*-value associated with the null hypothesis that all the experience indicators are equal is 0.0001. However, models with track-by-school effects are 66 percent smaller such that students with a teacher with 10 years of experience score only 0.01σ higher than a rookie teacher. The *p*-value associated with the null hypothesis that all the experience indicators are equal is 0.13— suggesting a weak association between teacher experience and English test scores within tracks. Approximately 60 percent of the variation in teacher experience occurs within tracks for both subjects. Accordingly, a lack of variability in experience within tracks *cannot* explain the attenuated effects for English— indicating that accounting for tracks is important.

V.3 Main Results

I now analyze the effects of accounting for tracks on the estimated variability of teacher effects. In Table 5, I present the estimated variability 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 all individual-student-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. Note that estimated standard deviations and covariances for models that include school-track fixed effects in the first stage have been inflated by the degree of freedom adjustment to account for the number of teachers in the school-track.

I start out discussing the variance subtraction results for algebra teachers. 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

0.18σ (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 classrooms is about 0.205σ ─ indicating that there is more variation in teacher performance across classrooms for the same teacher than there is across teachers. Because each teacher estimate is based on an average 3.6 classrooms, this implies that approximately $0.205/\sqrt{3.6}=0.109\sigma$ can be attributed to transitory variation (i.e. due to classroom 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 0.1367σ. In models that include school-track fixed effects the standard deviation of the raw teacher fixed effects falls to 0.1475σ and the adjusted standard deviation is 0.1069σ. This change is significant because the estimated variance without track effects is 1.64 times as large as that with track effects included (the estimated standard deviation without track effects is 1.27 times as large as that with track effects included). This indicates that selection to tracks or track specific treatments are important sources of variability across classrooms that should not be attributed to teachers. In the preferred model with school-track fixed effects and school-by-year effects (model 5) the standard deviation of the raw teacher fixed effects falls to 0.1237σ and the adjusted standard deviation of the teacher effects falls by to 0.0775σ.

The covariance-based estimates for algebra teachers are similar to those obtained using variance subtraction method— indicating that the main results are robust to the methodology used. The estimates indicate that the covariance of mean residuals across classrooms for the same teacher is 1.5 times larger without tracks that when track effects are accounted for— underscoring the importance of accounting for tracks. Models with school and year fixed effects and both student and peer covariates (model 3) yields a covariance across Algebra classrooms for the same teacher of 0.0148, yielding an estimated standard deviation of persistent teacher quality of 0.1215σ. This is similar to estimates from the variance subtraction method and is in-line with existing estimates in the literature. Adding additional controls for track effects (model 4) yields an estimated standard deviation of persistent teacher quality of 0.0998σ. The preferred model that includes track-school and school-year effects yields an estimated standard deviation of true teacher quality of 0.0677σ ─ similar to the estimates obtained using the variance subtraction approach. For all models one rejects the null hypothesis of zero covariance across classrooms for the same teacher at the 1 percent level. The 95% percent confidence interval for the preferred model ranges from 0.0514σ

to 0.0809σ. That is, going from an average teacher to one at the 85th percentile of the quality distribution increases algebra test scores by between 0.0514σ and 0.0809σ. Note that this covariance-based confidence interval includes the variance subtraction estimate. This confidence interval is informative because the 95% upper-bound is lower than many of the point estimates of existing studies (and lower than point estimates that do not account for school-track effects).

Now I discuss results for English teachers. In the variance subtraction model that only includes lagged test scores and school fixed effects (model 6), the standard deviation of the raw estimated teacher effects is 0.1025σ, and the adjusted estimate is 0.0713σ. In models with additional student controls (model 7), the standard deviation of the raw estimated teacher effects falls to 0.088σ and the adjusted estimate falls by 23 percent to 0.0579σ. Adding the characteristics of the peers in the classroom reduces the adjusted standard deviation slightly to 0.0514σ. Adding track-school effects, the adjusted standard deviation of teacher effects falls to 0.0405σ. This change is significant because it represents a 38 percent reduction in the estimated variance when track effects are included. In the preferred variance subtraction model (model 10) after adjusting these estimates for transitory variation and the degrees of freedom, the implied standard deviation of teacher effects falls by 24 percent to 0.0308σ. As expected, the biased F-tests reject the null hypothesis of no teacher effects at the 0.001 percent level.

The covariance-based results for English teachers are qualitatively similar to the variance subtraction results and reinforce to importance of accounting for track-level variability. As with the variance subtraction method, adding controls decreases the estimated variability of teacher effects. Adding student characteristics (other than a single lag of test scores) reduces the covariance across classrooms by about 38 percent, adding mean characteristics of classroom peers reduces the covariance by about 22 percent, adding track-school effects reduces the covariance by a further 40 percent, and adding school-year effects reduce the estimated variance by a further 62 percent. In the preferred covariance model (model 10) the estimated standard deviation of teacher effects is 0.0193σ (similar to the variance subtraction estimate). Unlike the biased *F*-test, the covariance-based test in the preferred models that includes track effects and school-year effects yields a *p*-values of 0.219. That is, while the widely used *F*-test erroneously rejects the null hypothesis of no teacher effects at the 0.001 percent level, the unbiased covariance based test cannot reject at the 10 percent level. This indicates that while there *may be* persistent English teacher effects, the data do not support this conclusion as strongly as previous studies might indicate. The confidence

interval for the standard deviation of English teacher effects is informative. Models that include school-track fixed effects and school-year effects yield a 95% percent confidence interval for the standard deviation between 0σ (no effect) to 0.0312σ (modest variability). This covariance-based confidence interval includes the variance subtraction estimate. The additional information provided by the covariance based results (i.e. unbiased *p*-values and confidence intervals) underscore the fact that excess variability that can be attributed to individual teachers is not necessarily indicative of persistent teacher quality differences. Given that much policy is predicated on the idea that a sizable amount of teacher quality is persistent over time, the fact that the 95% confidence interval lower bound is small for algebra (0.0514σ) and 0σ for English is important and policy relevant.

V.4 *Alternate explanations for the reduction in teacher effect variability*

Is this reduction in the variance of teacher effects due to removing single-classroom teachers?

Because the covariance estimates can only be estimated for teachers with multiple classrooms, it is possible that the covariance estimates yield a low variance because teachers with bad outcomes might leave the profession after teaching one class (and are not included in the covariance sample). To test for this, I created the variance subtraction estimates for only teachers with multiple classrooms to see if this was much smaller than the estimates that included all teachers. The estimated standard deviations are 5 percent smaller for Algebra and 7 percent smaller for English— suggesting that this does not drive the results. The fact that accounting for tracks reduced the variability of teacher effects in both the covariance and the variance subtraction methods indicates that the main results are not driven by any one individual estimation strategy.

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

While it is clear that not controlling for track effects (and attributing all track-level variability to individual teachers) may lead one to overstate the variability of teacher effects, it is also possible that removing track effects (attributing none of the track level variability to individual teachers) may lead one to understate the variability of teacher effects. This problem is likely most severe when tracks are observed with a small number of teachers. That is, unless there is severe sorting of teachers to tracks, it is much more likely that a high track-level mean is due to the track rather than teachers when it is based on several teachers (whose average effect should be close to the unconditional mean of teacher quality) than if it were based on only a handful of teachers.

Because many teachers are observed in multiple tracks, one can identify teacher and track effects simultaneously and let the data tell us whether the variation should be attributed to teachers or tracks. The logic is as follows; If teacher A has very good outcomes in track 1 while teacher B has poor outcomes in track 2, there is no way to know if the better outcomes for teacher A in track 1 are due to teacher A being a better teacher than teacher B or track 1 being a higher value-added track than track 2. However, if teachers A and B both have better outcomes in track 1 than in track 2, one can infer that track 1 is a high value-added track. Conversely if teacher A is better than teacher B in both tracks 1 and 2, one can infer that teacher A is a better teacher than teacher B. I follow Jackson (forthcoming) and exploit the within-teacher variation in tracks to separately identify track effects from teacher effects in a Restricted Maximum Likelihood (REML) model. First I estimate [5] below.

[5] $$Y_{igcjy} = A_{iy\text{-}1}\delta + X_i\beta + X^*_{jc}\pi + \theta_{sy} + \varepsilon^*_{ijcgy}$$

This model includes student and peer covariates and the school-by-year effects. The residuals from this regression $e_{ijgcy} \equiv \theta_g + \theta_j + \theta_{jc} + \varepsilon_{ijgcy}$ include the track effect, the teacher effect, the classroom error, and the random student level error. I then decompose the residual to estimate the true variance of the track, teacher, and classroom effects by maximum likelihood under the covariance structure described in [6] below.[13]

[6] $$Cov\begin{bmatrix}\theta_j \\ \theta_{jc} \\ \theta_g\end{bmatrix} = \begin{bmatrix}\sigma^2_{\theta j} I_J & 0 & 0 \\ 0 & \sigma^2_{\theta c} I_c & 0 \\ 0 & 0 & \sigma^2_{\theta g} I_G\end{bmatrix}.$$

By exploiting within-teacher variation in tracks, within-track variation in teachers, and modeling the finite-sample variability, the model apportions some of the track-level variability to the teacher and some to the track in the way that is most consistent with both the distributional assumptions and the data. Intuitively, if there is relatively little within-teacher variation in performance across tracks then most of the observed variation across tracks will be assigned to teachers (and *vice versa* if there is relatively little within-track variation in teacher quality). Also, the estimator accounts for finite-sample variability and apportions less (more) variability to track effects when track estimates are based on a smaller (greater) number of teachers.[14] When track,

[13] This is estimated in two steps because computing power constraints preclude estimating a model with thousands of track effects, thousands of teacher effects, and thousands of school-year effects simultaneously.

[14] The identifying assumption is that *realized* track and teacher effects (in a given dataset) come from a data-generating process in which they are uncorrelated, so that any correlation between effects is due to finite-sample

teacher, and classroom effects are estimated simultaneously (and allowed to compete for explanatory power), the estimated standard deviations are very similar to, and statistically indistinguishable from, both sets of estimates in Table 5. Specifically, the estimated standard deviation of Algebra teacher effects is 0.0680 and that for English teachers is 0.0245. Because the REML estimator relies on different identifying assumptions than the other models, the similarity across all models suggest that removing all track-level variability (which *could* have removed real variability in teacher quality) in the first stage is not a likely explanation for the modest estimated variability of teacher effects.

V.5 How predictive is estimated value-added of teachers' future performance?

To gauge the extent to which value-added estimates can predict teacher 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 [4] using 2005 through 2007 data, compute teacher value added based on residuals, standardize the estimates to be mean zero unit variance, and then estimate equation [7] on data from 2008 through 2010 where $z_{\hat{\theta}_j}$ is the estimated (pre sample) standardized 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 0.032σ. 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 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 19 percent of the the out-of-sample predictive power as those at elementary school. To further ensure that the lack of predictive power is not due to "over controlling", I estimate the same

variability. Under this condition, *after accounting for finite-sample variability*, the within-teacher variability in track effects is a reliable measure of the variability of track effects and the within-track variability in teachers is a reasonable measure of variability of teacher effects. As such, while the RMEL estimator is less likely to wrongly attribute teacher variation to tracks than the other models, if there is extreme sorting of teachers to tracks, the REML estimates may also be biased.

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* 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 column 7 show that a one standard deviation increase in estimated pre-sample value added raises English test scores by 0.0127σ. This effect is statistically significant at the 1 percent level. Similar to Algebra, the small magnitude of this effect is notable because this is about 20 percent of the size of out-of-sample predictive ability of estimated value-added for English teachers at the elementary-school level.

V.6 Are These Out of Sample Effects Driven by Student Sorting Within Tracks?

To assuage concerns that these out-of-sample effects are driven by student selection to teachers (i.e. teachers with high historical value added being assigned to students with higher ability), I determine if teacher value-added is correlated with *observable* student characteristics. To do this I compute 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.[15] I then run a regression of standardized teacher effects from 2005-2007 data on predicted outcomes from 2008-2010. With positive (or negative) assortative matching the coefficient on teacher value added would be positive (or negative). If there is little systematic sorting of students to teachers, the coefficient on pre-sample teacher value-added will be zero. The results are in the lower panel of Table 6. Historical teacher value added has no statistically or economically significant relationship with predicted Algebra test scores (despite having effects on actual test scores) in all models. For English, in models that do not include track effects, there is some evidence of positive assortative matching to teachers. However, in models that include school-track effects, historical teacher value added has no statistically or economically significant relationship with predicted English test scores (despite having effects on actual test scores). This suggests that while students may select to tracks, they *do not* select to individual teachers within tracks based on observables.

[15] 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.

Readers may also worry about selection to teachers based on *unobservables*. To test for student selection to teachers within school-track-years on *unobservables*, 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 equations [8] and [9] below separately on the data from 2008 through 2010 where $z_{\hat{\theta}_j}$ is the standardized estimated (pre sample) value added of teacher j, $\bar{z}_{\hat{\theta}_j}$ is the mean standardized 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 ψ_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 0.0277, and using changes in aggregate track level mean teacher quality (using only variation across years) yields a point estimate of 0.03596 (column 4). This pattern is robust to including school by year effects which yields a point estimate of 0.02786 (column 5) — suggesting that the estimated algebra teacher effects are real and

are not driven by selection to teachers within tracks. For English, the point estimates within school-track years is 0.01253 (column 8) and those for models that are robust to selection are similar at 0.0153 (columns 9) and 0.0182 (column 10). The results indicate that the estimated teacher effects are real, and are not driven by student selection to teachers across cohorts or student selection to teachers within school-track-years in either observable or unobserved dimensions.

VI Conclusions

Despite mounting evidence that elementary-school teachers have large effects on student test scores, much less is known about the effect of high-school teachers. I argue that in a high-school setting, (a) student selection to tracks may lead to biases in teacher value-added that cannot be accounted for with controls for past test scores or the covariates used in standard value-added models, and (b) 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 omitted variables bias due to "track treatment effects". These biases create additional 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 classroom-level disturbances and I propose an unbiased statistical test. Also, I propose a method for computing confidence intervals for the variability in teacher effects.

Using methods that account for track effects yields estimated teacher effects that are noticeably smaller than those obtained when track effects are not accounted for. A one standard deviation increase in algebra and English teacher quality is associated with 0.0677σ and 0.0193σ higher test scores, respectively. I demonstrate how using the *F*-test to test for the existence of English teacher effects yielded biased inference in previous studies. Using the proposed method, I find that the 95% confidence interval for the standard deviation of teacher effects in algebra is (0.0514 ; 0.081) while that for English teachers includes zero and is (0.0000 ; 0.0312). This range of plausible estimates is lower than point estimates found in the few existing studies of high school teachers, suggesting that accounting for track-level variability is important.

Even though this paper is focused on biases due to track effects in high school, it highlights the broader point that in all education contexts, *even with no differences in student ability across teachers,* teacher value added will be biased if there are unobserved "treatments" that differ systematically across teachers. At the elementary level, this could occur if some teachers are

always assigned to noisy classrooms (which lowers test scores), always teach early morning classes (which lowers test scores), or always teach the students who participate in the spelling bee (which increases test scores). This paper highlights that if one aims to accurately measure teacher quality, it is important to account for *all* systematic difference across teachers (and not just student selection).

I also investigate how estimated value-added predicts teachers' future performance. The results suggest that the scope for using value added in personnel decisions in high school might be limited. A good 9th grade teacher (85 percentile) based on previous performance raises current scores by about 0.028σ and 0.015σ more than an average teacher in algebra and English, respectively. 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 0.028*0.1=0.0028σ in algebra and by 0.018*0.1=0.0018σ 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 0.028*0.5=0.014σ (1.4 percent of a standard deviation) in algebra and by 0.018*0.5=0.009σ (0.9 of a percent of a standard deviation) in English. These calculations suggest that even with large changes to the teacher quality distribution, one cannot expect 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 five times smaller in high school than in elementary school.

There are numerous possible explanations for the differences between the two contexts: students may be more malleable at younger ages; younger children may exhibit more externalizing behaviors so that classroom management skills lead to larger differences in primary school outcomes; the match between the teacher and the student may be more important for adolescents such that teacher effectiveness changes substantially from year to year with different groups of students; secondary school course content may be less sensitive to teacher quality; and the testing instruments used in primary and secondary school may be differentially responsive to those skills that teachers affect. Given all these possible explanations for differences, these findings underscore the importance of not generalizing teacher effects obtained in one context across all grade levels and subjects. It is also important to note that, despite modest test-score effects, high-school teachers may have meaningful effects on important outcomes that are not well-captured by test

scores such as self-esteem, motivation, and aspirations. Jackson (2012) suggests that this is indeed the case.

In sum, this paper highlights some limitations of commonly used methodical tools in the literature, presents new tests that overcome these limitations, and proposes a method for computing confidence bounds for the variability of teacher effects. The results based on these new approaches indicate that one should avoid the temptation to use studies based on elementary-school teachers to make inferences about teachers in general, and demonstrates that the potential gains of using value added in personnel decisions in high school may be small. Overall, this paper speaks to the importance of using empirical methodologies that are appropriate to the specific context.

References

1. Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25: 95-135.

2. Balfanz, Robert, and Nettie Legters. 2004. Locating the Dropout Crisis: Which High Schools Produce the Nation's Dropouts? Where are they Located? Who Attends Them? *Center for Research on the Education of Students Placed at Risk Technical Report 70, Johns Hopkins University.*

3. Booker, Kevin., Tim. Sass, Brian. Gill, and Ron Zimmer. 2008. Going beyond test scores: Evaluating charter school impact on educational attainment in Chicago and Florida. *Santa Monica, CA: RAND Corporation.*

4. Chetty, Raj, John Friedman, and Jonah Rockoff. 2011. The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. *unpublished manuscript.*

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

6. Deming, Dave., Justine Hastings, Tom Kane, and Douglas Staiger. 2011. School Choice, School Quality and Academic Achievement. *unpublished mimeo.*

7. Frank, Kenneth A., Chandra Muller, Kathryn Schiller, Catherine Riegle-Crumb, Anna Strassmann Mueller, Robert Crosnoe, and Jennifer Pearson. 2008. The Social Dynamics of Mathematics. *American Journal of Sociology* 113 (6).

8. Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. Identifying E¤ective Teachers Using Performance on the Job. *Hamilton Project Discussion Paper 2006-01.*

9. Hanushek, Eric A. 2009. Teacher Deselection. In *Creating a New Teaching Profession*, by Dan Goldhaber and Jane Hannaway, 165-180. Washington, DC: Urban Institute Press.

10. Harris, Douglas N., and Andrew Anderson. 2013. Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence from Middle School Teachers. *Working Paper Tulane University.*

11. Heckman, James, and Paul. A. LaFontaine. 2010. The American High School Graduation Rate: Trends and Levels. *The Review of Economics and Statistics* 92 (May): 244-262.

12. Jackson, C. Kirabo. Forthcoming. Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics.*

13. Jackson, C. Kirabo. 2012. Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina. *National Bureau of Economics Reserch WP # 18624.*

14. Jackson, C. Kirabo. 2012. School Competition and Teacher Quality: Evidence from Charter School Entry in North Carolina. *Journal of Public Economics* 96 (5-6): 431–448.

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

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

17. Jacob, Brian A., and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics* 26 (1): 101–36.

18. Kane, Thomas, and Douglas Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER working paper 14607.*

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

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

21. Koedel, Cory, and Julian Betts. 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.*

22. Lochner, Lance, and Enrico Moretti. 2004. The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *American Economic Review,* 94 (1): 155-189.

23. Lucas, Samuel R., and Mark Berends. 2002. Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking. *Sociology of Education* 75 (4): 328-348.

24. Pleis, J.R., J.W. Lucas, and B.W. Ward. 2009. Summary Health Statistics for U.S. Adults: National Health Interview Survey, 2008. *Vital Health Stat* (National Center for Health Statistics) 10.

25. Protik, Ali, Elias Walsh, Alexandra Resch, Eric Isenberg, and Emma Kopa. 2013. Does Tracking of Students Bias Value-Added Estimates for Teachers? *Mathematica Working Paper* .

26. Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, Schools, and Academic Achievement. *Econometrica* 73 (2): 417-458.

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

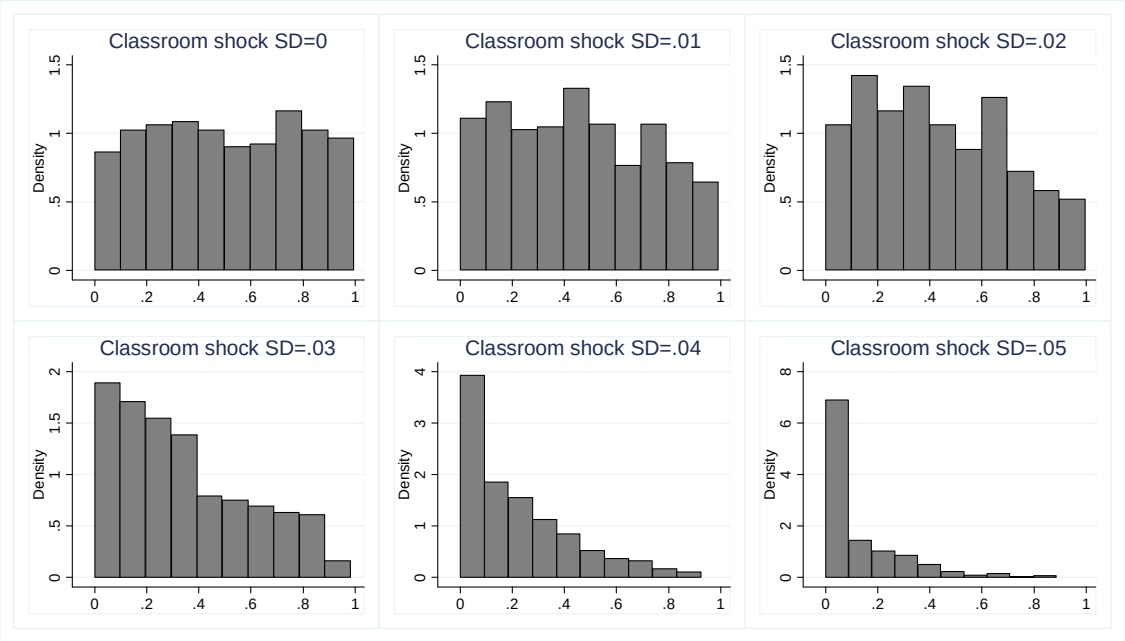
28. Rothstein, Jesse. 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics.*

29. Sadker, David M., and Karen Zittleman. 2006. *Teachers, Schools and Society: A Brief Introduction to Education.* McGraw-Hill.

30. Slater, Helen., Neil. M. Davies, and Simon Burgess. 2011. Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England. *Oxford Bulletin of Economics and Statistics.*

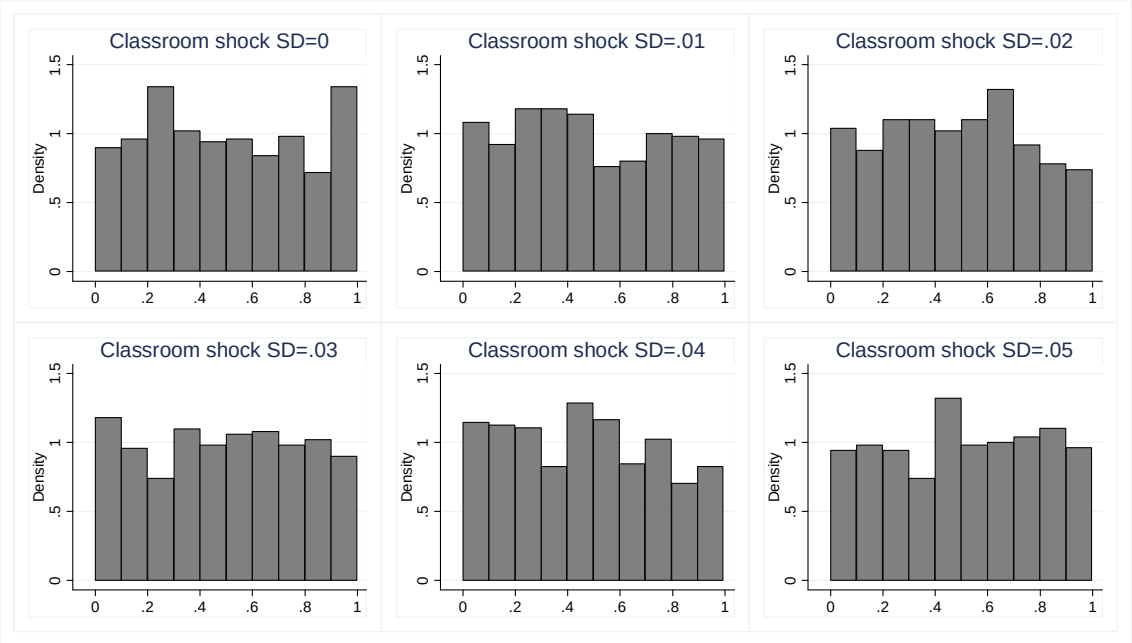
Tables and Figures

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



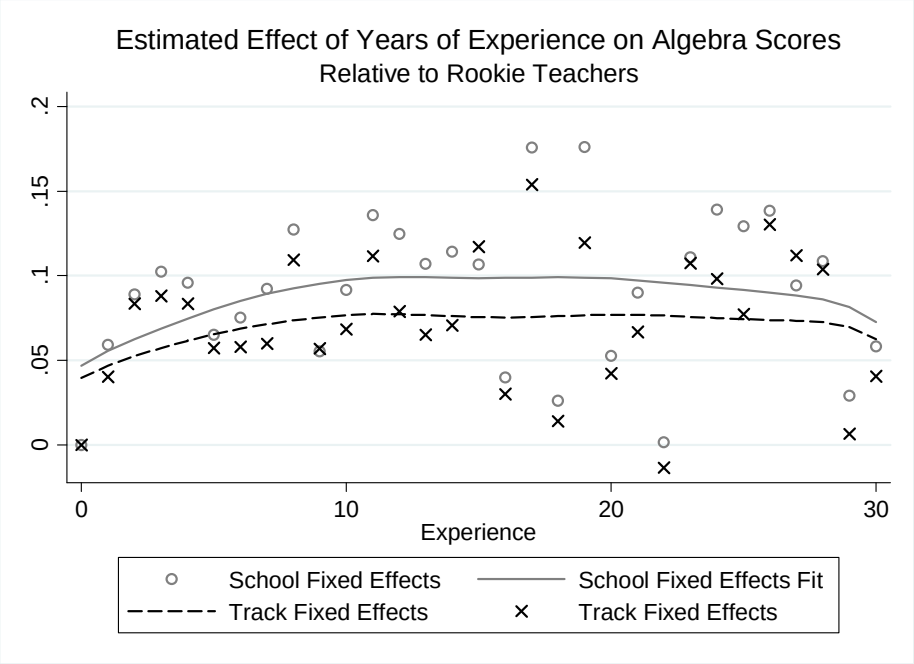
This figure plots the distribution of *p*-values of the hypothesis of no teacher effects when the null is true based on the F-tests under different standard deviation of classroom level errors. The F-tests includes a degree of freedom adjustment for clustering at the classroom level.

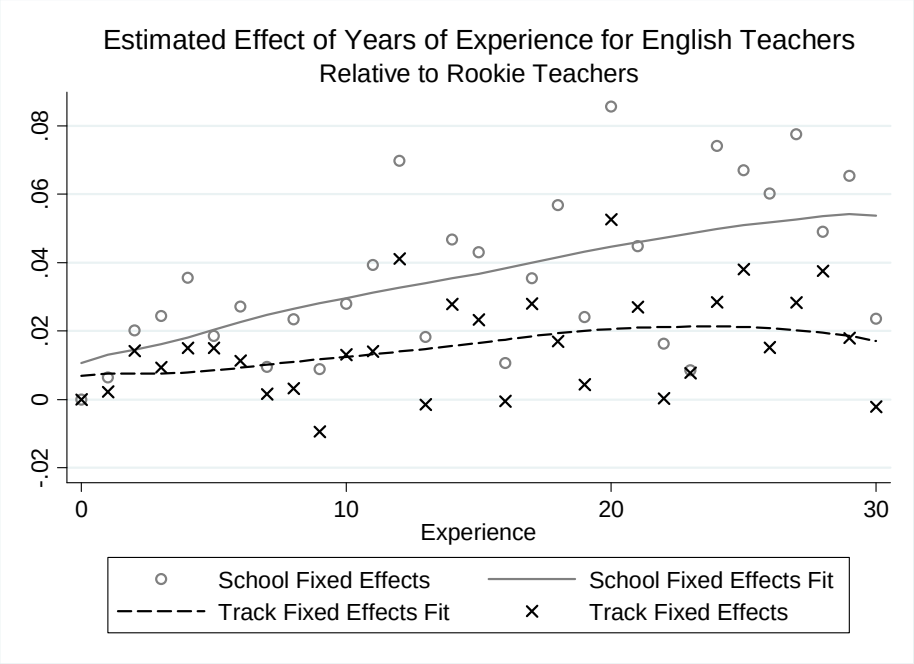
Figure 2: *Covariance Test (based on simulated data with no teacher effects)*



This figure plots the distribution of *p*-values of the hypothesis of no teacher effects when the null is true based on the covariance test under different standard deviation of classroom level errors.

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 school fixed effects, and track-by-school fixed effects.

Table 1: *Summary Statistics of Student Data*

Variable	Mean	SD	SD within school-tracks	SD within schools
Math z-score 7th grade	-0.003	(0.944)	(0.689)	(0.910)
Reading z-score 7th grade	-0.013	(0.961)	(0.682)	(0.904)
Math z-score 8th grade	0.091	(0.944)	(0.600)	(0.878)
Reading z-score 8th grade	0.073	(0.941)	(0.678)	(0.891)
Male	0.510	(0.50)	(0.482)	(0.498)
Black	0.288	(0.453)	(0.375)	(0.399)
Hispanic	0.075	(0.263)	(0.245)	(0.256)
White	0.579	(0.494)	(0.404)	(0.432)
Asian	0.020	(0.141)	(0.133)	(0.138)
Parental education: Some High-school	0.075	(0.263)	(0.25)	(0.259)
Parental education: High-school Grad	0.400	(0.49)	(0.454)	(0.474)
Parental education: Trade School Grad	0.018	(0.132)	(0.129)	(0.132)
Parental education: Community College Grad	0.133	(0.339)	(0.327)	(0.335)
Parental education: Four-year College Grad	0.205	(0.404)	(0.376)	(0.394)
Parental education: Graduate School Grad	0.064	(0.245)	(0.225)	(0.237)
Number of Honors classes	0.880	(1.323)	(0.575)	(1.163)
Algebra I z-Score (9th grade)	0.063	(0.976)	(0.775)	(0.889)
English I z-Score (9th grade)	0.033	(0.957)	(0.670)	(0.906)
Observations			348547	

Notes: These summary statistics are based on student who took the English I exam. Incoming math scores and reading scores are standardized to be mean zero unit variance. About 10 percent of students do not have parental education data so that the missing category is “missing parental education”.

Table 2: *Most common academic courses*

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

Table 3: *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 4: *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	
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	

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.
Standard errors in brackets are clustered at the teacher level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: *Variability of Teacher Effects*

	Algebra				
	1	2	3	4*	5*
	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
SD of teacher level residuals	0.1803	0.1745	0.1741	0.1475	0.1237
F-test: Prob(all teacher level means equal)	0.0000	0.0000	0.0000	0.0000	0.0000
SD of classroom effects (within teachers)	0.2097	0.2042	0.2044	0.1928	0.1830
SD of mean of transitory variability	0.1105	0.1076	0.1077	0.1016	0.0964
Implied variance of persistent teacher effects	0.0203	0.0189	0.0187	0.0114	0.0060
Implied true SD of persistent teacher effects	0.1425	0.1373	0.1367	0.1069	0.0775
Covariance across classrooms	0.0155	0.0150	0.0148	0.0091	0.0046
SD of the sample covariance	0.0010	0.0009	0.0009	0.0009	0.0010
t-stat	16.2428	16.4413	15.6864	9.5550	4.7086
t-test: Prob(covariance=0)	0.0000	0.0000	0.0000	0.0000	0.0000
SD of Teacher effects (covariance method)	0.1243	0.1223	0.1215	0.0998	0.0677
95% CI Lower bound	0.1164	0.1146	0.1135	0.0846	0.0514
95% CI Upper bound	0.1318	0.1295	0.1290	0.1046	0.0809
	English				
	6	7	8	9*	10*
SD of teacher-level residuals	0.1025	0.0883	0.0835	0.0741	0.0646
F-test: Prob(all teacher-level means equal)	0.0000	0.0000	0.0000	0.0000	0.0000
SD of classroom effects (within teachers)	0.1680	0.1520	0.1501	0.1415	0.1295
SD of mean of transitory variability	0.0737	0.0667	0.0658	0.0621	0.0568
Implied variance of persistent teacher effects	0.0051	0.0034	0.0026	0.0016	0.0009
Implied true SD of persistent teacher effects	0.0713	0.0579	0.0514	0.0405	0.0308
Covariance across classrooms	0.00337	0.00208	0.00162	0.00098	0.00037
SD of the sample covariance	0.0005	0.0004	0.0003	0.0003	0.0003
t-stat	6.8449	5.7766	5.3685	3.2667	1.2284
t-test: Prob(covariance=0)	0.0000	0.0000	0.0000	0.001	0.2193
SD of Teacher effects (covariance method)	0.0581	0.0456	0.0402	0.0313	0.0193
95% CI Lower bound	0.0489	0.0369	0.0319	0.0194	0.0000
95% CI Upper bound	0.0660	0.0529	0.0471	0.0397	0.0312

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, and gender. Peer covariates are the classroom-level means of all the student-level covariates. There are 5.4 classes per English teacher and 3.6 classes per algebra teacher.

* The reported covariance or SD of teacher effects includes a degree of freedom adjustment for the number of teachers in a track in models that condition on school track in the first stage (models 4, 5, 9, and 10).

Table 6: *Out of Sample Predictions*

	Algebra scores					English Scores				
	1	2	3	4	5	6	7	8	9	10
	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	
Standardized Teacher Effect (2005-7)	0.03227***	0.03201***	0.02770***			0.01111***	0.01277***	0.01253***		
	[0.006]	[0.006]	[0.006]			[0.002]	[0.002]	[0.002]		
Mean Standardized Teacher Effect (2005-7)				0.03596***	0.02786**				0.01534***	0.01827***
				[0.010]	[0.013]				[0.005]	[0.006]
	Predicted Algebra scores					Predicted English Scores				
Standardized Teacher Effect (2005-7)	-0.00317	-0.00415	0.00438			0.02030**	0.00394	0.00372		
	[0.007]	[0.004]	[0.005]			[0.008]	[0.003]	[0.003]		
Mean Standardized Teacher Effect (2005-7)				-0.00439	0.01147				-0.00528	-0.005
				[0.003]	[0.011]				[0.006]	[0.006]
Observations	44,729	44,745	44,745	67,120	67,100	105,555	105,591	105,591	130,949	130,907
Year Effects	Y	-	-	Y	-	Y	-	-	Y	-
School Effects	Y	-	-	-	-	Y	-	-	-	-
Track-School-Effects	N	Y	-	Y	Y	N	Y	-	Y	Y
School-Year-Effects	N	Y	-	N	Y	N	Y	-	N	Y
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

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, and gender. Peer covariates are the classroom-level means of all the student level covariates.

Standard errors in brackets are clustered at the teacher level. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix

Appendix A

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: *Distribution of student Observations by the Distribution of Number of Teachers in Each School-Track Cell with The Implied required Degree of Freedom Adjustment*

Teachers per track	Algebra	English	DOF adjustment
1	0.032	0.039	-
2	0.074	0.073	2.000
3	0.069	0.080	1.500
4	0.073	0.081	1.333
5	0.080	0.085	1.250
6	0.088	0.076	1.200
7	0.081	0.080	1.167
8	0.081	0.073	1.143
9	0.056	0.053	1.125
10	0.065	0.053	1.111
11	0.059	0.048	1.100
12	0.040	0.036	1.091
13	0.038	0.040	1.083
14	0.029	0.030	1.077
15	0.023	0.027	1.071
16	0.021	0.018	1.067
17	0.011	0.012	1.063
18	0.009	0.015	1.059
19	0.004	0.010	1.056
20+	0.069	0.072	1.042
Average Weight	1.239	1.244	

Table A3: *Distribution of the number of classes per teacher and per teacher year*

Number of classes	Proportion of teacher observations		Proportion of teacher-year observations	
	Algebra	English	Algebra	English
1	28.72	24.5	43.83	36.33
2	20.44	14.84	31.42	26.81
3	14.32	10.7	14.37	18.91
4	9.92	8.79	6.97	10.08
5	6.89	6.19	2.39	5.18
6	4.31	5.61	0.83	2.13
7	4.03	4.68	0.15	0.33
8	2.58	3.7	0.05	0.17
9	2.3	3.21	0	0.03
10	1.67	2.95	0	0.03
11	1.13	2.77	0	0.02
12	0.88	2.23	0	0
13	0.45	1.72	0	0
14	0.43	1.3	0	0
15	0.34	1.16	0	0
16	0.31	0.98	0	0
17	0.4	0.84	0	0
18	0.26	0.6	0	0
19	0.11	0.6	0	0
20+	0.53	2.61	0	0

Table A4: *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 B

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 (5 percent). 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 (11 percent). 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 (17 percent). 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 (26 percent). Matched classes are set aside.
(5) Repeat step 4 until no more high quality matches can be made (6 percent).

This procedure leads to a matching of approximately 65 percent of classrooms in the End of Course files. This corresponds to 57 percent of all Algebra student observations and 69 percent of all English student observations. The demographic characteristics of students in matched classroom are very similar to those in unmatched classrooms (see table below). However, unmatched classrooms appear to have somewhat higher incoming student test scores that the full sample. 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.

Table B1: Characteristics of Matched and Unmatched Students

Variable	Matched	All students
Male	0.510	0.511
Black	0.288	0.295
Hispanic	0.075	0.073
White	0.579	0.580
Asian	0.020	0.020
Parental education: Some High-school	0.075	0.076
Parental education: High-school Grad	0.399	0.380
Parental education: Trade School Grad	0.018	0.019
Parental education: Community College Grad	0.133	0.129
Parental education: Four-year College Grad	0.205	0.209
Parental education: Graduate School Grad	0.064	0.069
Math z-score 7th grade	-0.003	0.007
Reading z-score 7th grade	-0.013	0.013
Math z-score 8th grade	0.091	0.124
Reading z-score 8th grade	0.073	0.133

Deriving the degree of Freedom Adjustment for the Size of Tracks

The track specific fixed effect estimated in the first stage is the mean residual test scores among student in the track. Assuming no track-specific treatment we have [a] below, where w_{ig}, w_{cjg}, and w_{jg}, weight student errors, classroom errors, and teacher effects respectively, by the number of observations associated with them in track g.

[a] $$\hat{\theta}_g = \sum_{i=1} w_{ig}\varepsilon_{ijcgy} + \sum_{cj=1} w_{cjg}\mu_{cj} + \sum_{j=1} w_{jg}\theta_j$$

Removing this track-specific mean from the raw classroom level residual yields

[b] $$\bar{e}^*_{jc} = \left(\theta_j - \sum_{j=1} w_{jg}\theta_j\right) + \left(\mu_{cj} - \sum_{cj=1} w_{cjg}\mu_{cj}\right) + \left(\varepsilon_{ijcgy} - \sum_{i=1} w_{ig}\varepsilon_{ijcgy}\right).$$

Because the student level and classroom level errors are random, taking the covariance of these classroom effects for the same teacher yields

[c] $$Cov(\bar{e}^*_{jc}, \bar{e}^*_{jc'}) = Cov\left(\left(\theta_j - \sum_{j=1} w_{jg}\theta_j\right), \left(\theta_j - \sum_{j=1} w_{jg}\theta_j\right)\right) = Var\left(\theta_j - \sum_{j=1} w_{jg}\theta_j\right).$$

If teachers are randomly distributed across tracks, this simplifies to $Var(\theta_j)\left(1 - \sum \omega^2_{jg}\right)$. With equal weighting for teachers in a track this simplifies further to $Var(\theta_j)\left(1 - (1/R_g)\right)$, where R_g is the number of teachers in track g. This implies that while removal of a track specific mean does remove bias due to track-specific treatment, it may lead one to understate the covariance across classrooms by a factor of $\left(1 - (1/R)\right) < 1$, where R_g is the number of teachers per track.

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[1]. 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.

Estimating Efficient Teacher Fixed Effects

I follow the procedure outlined in Kane and Staiger (2008)[2] to compute efficient teacher fixed effects. This approach accounts for the fact that (1) teachers with larger classes will tend to have more precise estimates and (2) there are classroom level disturbances so that teachers with multiple classrooms will have more precise estimates. As before, I compute mean residuals from [4] for each classroom $\overline{e}^{*}_{cj} \equiv \theta_j + \mu_{cj} + \hat{\varepsilon}_c$ where $\hat{\varepsilon}_c$ is the mean individual student level error in classroom *jc* . Since the classroom error is randomly distributed, I use the covariance between the mean residuals of classrooms for the same teacher $\mathrm{cov}(\overline{e}^{*}_{cj}, \overline{e}^{*}_{c'j}) = \hat{\sigma}^2_{\theta_j}$ as an estimate of the variance of true teacher quality. I use the variance of the classroom demeaned residuals as an estimate of $\hat{\sigma}^2_{\varepsilon}$. Because the variance of the residuals is equal to the sum of the variances of the true teacher effects, the classroom effects, and the student errors, I compute the variance of the classroom errors σ^2_c by subtracting σ^2_{ε} and $\hat{\sigma}^2_{\theta_j}$ from the total variance of the residuals. For each teacher I compute [d], a weighted average of their mean classroom residuals, where classrooms with more students are more heavily weighted in proportion to their reliability.

[d] $$\hat{\theta}_j = \sum_{t=1}^{T_j} z_{jt} \cdot \frac{(1/(\sigma^2_c + (\sigma^2_{\varepsilon}/N_c))}{\sum_{t=1}^{T_j}(1/(\sigma^2_c + (\sigma^2_{\varepsilon}/N_c))}$$

Where N_c is the number of students in classroom *c*, and T_j is the total number of classrooms for teacher *j*. The correlation between this measure and the teacher-level mean residual is above 0.9 for both subjects.

[1] 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.
[2] A similar approach is also used in Jackson (2009, 2012).

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. The table below 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.[3] 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 B2: *Evidence of Students Sorting to Teachers based on Prior Test Scores*

	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	0.1530	0.1683	0.1434	0.1720	0.1788	0.1502
Full sorting across	0.0012	0.0012	0.0012	0.0017	0.0010	0.0014
Random assignment	0.6514	0.6360	0.4978	0.7445	0.7138	0.5645
Random assignment	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.

[3] 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.