2018

What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes

C. Kirabo Jackson

What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test-Score Outcomes[1]

C. Kirabo Jackson, 15 April 2017
Northwestern University, and National Bureau of Economic Research

I present a model in which teachers affect a variety of student outcomes through their influence on both cognitive and noncognitive skill. Empirically, I proxy for students' noncognitive skill using non-test-score behaviors. These behaviors include absences, suspensions, course grades, and on-time grade progression in 9th grade. Teachers have meaningful effects on both test scores and behaviors. However, teacher effects on test scores and those on behaviors are weakly correlated. Teacher effects on noncognitive proxy measures (i.e. behaviors) predict larger impacts on high-school completion and other longer-run outcomes than their effects on test-scores. Relative to using only test-score measures, using teacher effects on both test-score and noncognitive proxy measures more than doubles the variance of predictable teacher impacts on longer-run outcomes. (JEL I21, J00)

At the broadest level, a good teacher is one who teaches students the skills needed to be productive adults (Douglass 1958; Jackson, Rockoff, and Staiger 2014). Not every skill needed in adulthood is well-captured by performance on achievement tests. Indeed, a large body of research demonstrates that "noncognitive" skills not captured by standardized tests, such as adaptability, self-restraint, and motivation, are key determinants of adult outcomes.[2] Even so, economists have focused on test-score measures of teacher quality (referred to as test-score value-added) because they are often the best available measure of student skills.[3] However, good teachers may affect students much more broadly than through their impact on achievement test scores.

Chetty, Friedman, and Rockoff (2014b) show that teachers who improve test scores improve students' high-school completion, college-attendance, and earnings. While this finding shows the importance of measuring teacher's impacts on test scores, it does not show that impacts on test-scores provide a comprehensive measure of teacher quality. The literature on noncognitive skills provides reason to suspect that teachers may influence skills and behaviors that go undetected by test scores, but are nonetheless important for students' longer-run success. Because

1 I thank David Figlio, Jon Guryan, Simone Ispa-Landa, Clement Jackson, Shayna Silverstein, Mike Lovenheim, James Pustejovsky, Jonah Rockoff, Dave Deming, Jim Heckman, Alexey Makarin, Laia Navarro-Sola, and four anonymous reviewers for insightful comments and feedback. I also thank Kara Bonneau from the NCERDC. This research was supported by the Smith Richardson Foundation and the Carnegie Corporation on New York.
2 See Lindqvist and Vestman (2011), Heckman and Rubinstein (2001), Waddell (2006), and Borghans, Weel, and Weinberg (2008). Consistent with this, some interventions that have no effect on test scores have meaningful effects on long-term outcomes (Booker et al. 2011; Deming, 2009; Deming, 2011), and improved noncognitive skills explain the effect of some interventions (Fredriksson, Ockert, and Osterbeek 2013; Heckman, Pinto, and Savelyev 2013).
3 Having a teacher at the 85th versus the 15th percentile of the test-score value-added distribution is found to increase test-scores by between 8 and 20 percentile points (Kane and Staiger, 2008; Rivkin, Hanushek, and Kain, 2005).

districts seek to measure teacher quality for policy purposes, it is important to measure teacher effects on overall well-being and not *only* effects on those skills measured by standardized tests.

This paper estimates teacher effects on both test scores and measures of noncognitive skills.[4] I refer to a teacher's effect on a skill measure as value-added. I demonstrate that teachers have meaningful value-added on both test scores and noncognitive skill-measures in ninth grade. Surprisingly, test-score and noncognitive value-added for the same teacher are weakly correlated (r=0.15). I show that ninth-grade teachers who raise students' noncognitive skill-measures have important impacts on students' life chances. A one standard deviation increase in noncognitive value-added increases students' likelihood of graduating from high school by 1.47 percentage points, compared to only 0.12 percentage points for test-score value-added. This pattern of larger impacts of noncognitive value-added replicates across several high school outcomes, including grade progression, SAT-taking, 12th grade GPA, and intentions to attend a four-year college.

To motivate the empirical work, I follow Cunha and Heckman (2008) and extend the standard test-score value-added model that assumes unidimensional student ability. In this extended model, student outcomes are a function of both cognitive and noncognitive skills (Heckman, Stixrud, and Urzua 2006). I propose that one can measure teacher value-added on multiple ninth-grade skill measures, but focus on two; test scores and some other outcome. I show that, as long as the two skill measures do not reflect the same exact mix of student abilities, one can better predict teacher impacts on longer-run student outcomes using value-added on both skill measures than using test-score value-added alone. I then test this implication empirically.

I employ administrative data on all public school ninth-graders in North Carolina from 2005 through 2012. These data contain student scores on math and English exams linked to their subject teachers. To obtain measures of student skills in ninth grade that may not be well-captured by test scores, I follow a literature that uses behaviors as proxies for noncognitive skills.[5] To summarize these behaviors with a single variable, I use principal component analysis to create a weighted average of grades, on-time grade progression, absences, and suspensions. I refer to this

[4] Alexander, Entwisle, and Thompson (1987), Ehrenberg, Goldhaber, and Brewer (1995), Downey and Shana (2004), Jennings and DiPrete (2010), and Mihaly et al. (2013) find evidence that teachers have effect on non-test-score measures of student skills. Also, Koedel (2008) estimates high-school teacher effects on graduation.

[5] For example, Heckman, Stixrud, and Urzua (2006), Lleras (2008), Bertrand and Pan (2013), Kautz (2014), Heckman, Humphries and Varemendi (2016). In the same way that one infers that a student who scores higher on tests likely has higher cognitive skills than a student who does not, one can infer that a student who acts out, skips class, and does not hand in homework likely has lower noncognitive skills than a student who does not (Heckman and Kautz. 2012).

weighted average of ninth-grade behaviors as the “behavior index.” I use the behavior index to measure noncognitive skills *that may be missed by test scores*. However, I do not claim that this index is unrelated to cognitive skills, nor is this index a comprehensive measure of all noncognitive skills. Using value-added models, I estimate ninth-grade teacher effects on both test scores and behaviors. I then examine how teachers with high noncognitive value-added (i.e. those who improve behaviors) influence longer-run outcomes such as high-school completion, SAT-taking, and intended college going in ways unmeasured by their test-score value-added.[6]

Teachers in ninth grade have meaningful effects on both test scores and those behaviors that proxy for noncognitive skills. Teacher value-added on test scores and the behaviors are weakly correlated (r=0.15), and, conditional on test-score value-added, there is considerable variability in behaviors value-added. In models that predict high-school graduation using only test-score value-added, a one standard deviation increase in test-score value-added raises the likelihood of high-school graduation by 0.15 percentage points. However, when also including behaviors value-added, a one standard deviation increase in test-score value-added leads to 0.12 higher likelihood of graduation, and a one standard deviation increase in behaviors value-added leads to 1.47 percentage points higher likelihood of graduating from high school. These results suggest that (a) many teachers who raise test scores do not improve behaviors and vice versa and (b) behaviors value-added detects effects on important skills that are not detected by test-scores. Including both value-added measures more than doubles the predictable teacher-level variability in high-school graduation. Patterns are similar for dropout, SAT-taking, and college plans.

To address concerns of sorting and selection biases, all models include a rich set of covariates, and I present several empirical tests to show that the relationships presented can be interpreted causally. Moreover, I also show that these patterns are robust to using behavioral outcomes that cannot be driven by grade inflation or reporting biases (i.e. 10^{th} grade GPA).

The results support an idea that many believe to be true but that has not previously been shown – that teacher effects on test scores capture only a fraction of teacher effects on human capital. This underscores the need for evaluations that account for effects on both cognitive and noncognitive skills (Heckman 1999). Because some of the non-test-score outcomes used can be manipulated by teachers, using them directly for accountability or evaluation purposes is unwise.

[6] These longer-run outcomes are worthy of study because they include strong predictors of college-going, and high-school dropout is a strong predictor of crime, employment, and earnings.

However, I present some feasible policy uses. The results provide an explanation for why Chamberlain (2013) finds that test-score value-added may reflect less than one-fifth of the total effect of teachers. Also, consistent with Heckman, Pinto, and Savelyev (2013), teacher effects on proxies for noncognitive skills offers an explanation for why teacher test score effects fade over time (Jacob, Lefgren, and Sims 2010) despite having meaningful effects on long-run outcomes.

The remainder of this paper is organized as follows: section II describes the data. Section III presents the theoretical framework. Section IV presents the empirical framework. Section V analyzes short-run teacher effects. Section VI analyzes how short-run teacher effects predict longer-run teacher effects and discusses possible policy applications. Section VII concludes.

II Data and Relationships Between Variables

I seek to obtain estimates of the effect of ninth grade teachers on both test scores and proxies for noncognitive skills in ninth grade. I will then explore whether these estimated effects on ninth grade skill measures predict teacher impacts on longer-run outcomes. I use data on all public-school ninth grade students in North Carolina between 2005 and 2012 obtained from the North Carolina Education Research Data Center. The data include demographics, transcript data, test scores in grades seven through nine, and codes linking student test scores to the teacher who administered the test.[7] I focus on students who took English (English I) and math (algebra I, geometry, or algebra II) courses during ninth grade. Roughly 93 percent of all ninth graders take both English I and one of these math courses. To avoid any bias that would result from teachers influencing students' ninth grade repetition, I use only the first observation of ninth grade repeaters.[8] Summary statistics are presented in table 1.

These data cover 573,963 ninth grade students in 872 secondary schools, with 5,195 English teachers, and 6854 Math teachers. The gender split is roughly even. The sample is 58.8 percent white, 26.1 percent black, 7.2 percent Hispanic, and 2.1 percent Asian. Regarding the highest level of education obtained by either of the student's two parents, 46 percent had a high-school degree or less, 14.9 percent had a junior college or trade school degree, 29.4 percent had a four-year college degree or higher, and 9.5 percent are missing data on parental education. All test-score variables are standardized to be mean zero, unit variance, for the full population taking each

[7] I use an algorithm to ensure high quality matching of students to teachers. I detail this in Appendix A.
[8] Results that exclude ninth-grade repeaters entirely are essentially unchanged.

test during each testing year. Test scores in the sample are higher than average because the ninth graders successfully matched to their classroom teacher are slightly higher-achieving on average.[9]

Informed by studies that use behaviors as proxies for noncognitive skills not measured well by test scores (Lleras 2008; Bertrand and Pan 2013; Kautz and Zanoni 2014; Heckman, Humphries, and Veramendi 2016), I proxy for noncognitive skills using non-test-score behaviors available in the data: the log of the number of absences in ninth grade (plus 1), whether the student was suspended during ninth grade, the grade point average (based on all ninth-grade courses), and whether the student enrolled in tenth grade on time. These behaviors are strongly associated with well-known psychometric measures of noncognitive skills including the “big five” and grit.[10] Informed by Heckman, Stixrud, and Urzua (2006), I use a principal component model to create a single index of these behaviors. This index is a weighted average of the non-test-score outcomes, and is standardized to be mean zero and unit variance. I refer to this index as the *behavior index*.[11] The behavior index has a correlation of 0.56 with test scores. However, analysis of variance (ANOVA) reveals that about 75 percent of the variation in the behavior index is unrelated to test scores. As such, there is much variation in this index that is unrelated to test scores that may serve as a proxy for noncognitive skills that go largely unmeasured by standardized tests.[12]

The main longer-run outcomes analyzed are measures of high-school completion. Data on high-school dropout and graduation (through 2014) are linked to the 2005 through 2011 ninth grade cohorts. Graduation and dropout are measured for those in the public school system in North Carolina. Individuals who move out-of-state or to private school are neither graduates nor dropouts. As such, opposite effects observed on both outcomes cannot be due to changes in private school or out-of-state enrollment. While having both measures is valuable, high school dropout is

---

9 Also, test scores in seventh and eighth grades are higher than the average because (a) the sample is based on those higher achievers who remained in school through ninth grade, and (b) I use the most recent eighth- or seventh-grade score prior to ninth grade, which tends to be higher for repeaters.

10 Low agreeableness and high neuroticism are associated with more absences, externalizing behaviors, delinquency, and lower educational attainment (Lounsbury et al. 2004; Barbaranelli et al. 2003; John et al. 1994; Carneiro, Crawford, and Goodman 2007). High conscientiousness, persistence, grit, and self-regulation are associated with fewer absences and externalizing behaviors, higher grades, and on-time grade progression (Duckworth et al. 2007).

11 I estimated a principal component model on the behavioral outcomes. There is only one principal component (the first eigenvalue is 0.98 and the second is 0.010). I then computed the unbiased prediction of this sole underlying component using the Bartlett method. The predicted index equals 0.38(GPA) + 0.31(enrolled in tenth grade) – 0.15(suspended) – 0.21(log of 1 + absences). See Appendix B for correlations between the ninth-grade outcomes.

12 For example, GPA and test scores both measure some of the same academic cognitive skills. However, teachers base their grading on some combination of student product (exam scores, final reports, etc.), student process (effort, class behavior, punctuality, etc.) and student progress (Howley, Kusimo, and Parrott, 2000; Brookhart, 1993) so that grades reflect a combination of skills, only some of which may be measured by test scores.

notoriously difficult to measure (Tyler and Lofstrom 2009). As such, I focus analysis on the more reliable high-school graduation outcome. Roughly 4.3 percent of ninth graders are recorded as having subsequently dropped out of school, while about 82 percent graduated from high school.[13] The remaining 11 percent either transferred out of the North Carolina system or remained in school beyond the expected graduation year. Other longer-run outcome data include GPA at graduation, taking the SAT, and reported intentions to attend a four-year college upon graduation (2006 through 2011 cohorts). Roughly 48 percent of ninth graders took the SAT by 12th grade, and 35 percent intended to attend a four-year college.

To present suggestive evidence that these behaviors may proxy for skills not well-measured by test scores, I examine whether these behaviors (in ninth grade) predict the longer-run outcomes conditional on test scores in ninth grade (table 2). To remove the influence of socio-demographics, all models include controls for parental education, gender, ethnicity, English and math test scores, repeater status, absences, out-of-school suspension in seventh and eighth grade, GPA in eighth grade, and include indicator variables for each secondary school. Transcript data are only available in high school so that eighth grade GPA is only observed for high-school courses taken while in eighth grade (about 25% of students).[14] Appendix F shows that the main results are robust to excluding ninth-grade GPA as a skill measure and relying on the other behaviors for which the lags are observed for all students. Columns 1 and 2 show that higher test scores in ninth grade predict less dropout and more high-school graduation. Also, the non-test-score behaviors in ninth grade predict variation in these outcomes conditional on test scores. The coefficients on the individual behaviors all have the expected signs, and are statistically significant.

To facilitate an apples-to-apples comparison with the behavior index, I create a test-score index that is the average of ninth grade math and English scores. For both longer-run outcomes, increases in the behavior index are associated with sizeable improvements conditional on test-scores (columns 3 and 4). While a 1σ increase in the test-score index is associated with a 1.33 percentage point decrease in dropout, a 1σ increase in the behavior index is associated with a 5.24 percentage point decrease. Similarly, while a 1σ increase in the test-score index is associated with a 1.86 percentage point increase in high-school graduation, a 1σ increase in the behavior index is

---

13 These are verified dropouts. The low dropout rate reflects the fact that a dropout is often difficult to verify.

14 In regression models, those with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator that is equal to one for all such observations. All results are robust to excluding eighth-grade GPA.

associated with a 15.8 percentage point increase. Columns 5 through 8 present patterns for high-school GPA, SAT taking, and intentions to attend a four-year college. Across all the longer-run outcomes, increases in the behavior index are associated with large and statistically significant improvements, conditional on test-scores.[15] This suggests that teacher impacts on behaviors (a proxy for noncognitive skills) may be a good predictor of impacts on longer-run outcomes, above and beyond that predicted by their impacts on test scores. This is explored directly in section V.

III Theoretical Framework

The standard value-added model assumes that student ability is one-dimensional (see Todd and Wolpin 2003). Following Cunha and Heckman (2008) and Cunha, Heckman and Schennach, (2010), I extend this model so that student outcomes are functions of *both* cognitive and noncogntive abilities. In the model, teachers can improve skills that lead to improved longer-run outcomes but are not relfected in improved test scores. As such, teacher impacts on non-test-score outcomes can provide additional infomation (above and beyond that contained in teacher impacts on test scores) on the extent to which they improve longer-run otucomes. For expositional purposes, I refer to students' latent competencies as *abilities*, I refer to short-run student outcomes used to infer these competencies (such as test scores, course grades, etc.) as *skill measures*, and I refer to longer-run outcomes (such as high-school graduation and college going) as *outcomes*.

III.A Model Setup

Production of Student Skills: Prior to ninth grade, each student *i* has a stock of cognitive and noncognitive abilities described by vector $\nu_i = (\nu_{ci}, \nu_{ni})^T$, where the subscripts *c* and *n* denote the cognitive and noncognitive dimensions, respectively.[16] This stock reflects an initial endowment and the cumulative effect of all school and parental inputs on students' incoming abilities. Each ninth-grade teacher j has a positive quality vector $\omega_j = (\omega_{cj}, \omega_{nj})^T$ describing teacher j's capacity to increase each of the two dimensions of student ability during ninth grade. Each student has a matrix given by $D_i = \begin{bmatrix} D_{ci} & 0 \\ 0 & D_{ni} \end{bmatrix}$, that describes student *i*'s responsiveness to teacher quality in each dimension. The "effective" quality of teacher *j* for student *i* (ω_{ij}) is the

15 In Appendix C, I present similar patterns using nationally representative survey data, and I also present additional empirical patterns that validate the use of the behavior index as a proxy for noncognitive skills.

16 Students may possess many types of cognitive and non-cognitive skills. The key point is that the extension relaxes the assumption that students are either greater or lesser skilled, and permits the more realistic scenario in which students may be highly skilled in certain dimensions but deficient in other dimensions of skill.

student matrix D_i times the underlying quality vector of teacher *j* given by $\omega_{ij} = D_i \omega_j$.[17]

During ninth grade, students take classes in many subjects (i.e. math, English, sciences, social studies, etc.). The two-dimensional vector φ_{i-j} represents the contribution of the ninth grade teachers *other* than teacher *j* to the end-of-year ability of student *i*. Ability of student *i* at the end of ninth grade with teacher *j* is represented by the vector in [1].[18]

[1] $$\alpha_{ij} = \upsilon_i + \omega_{ij} + \varphi_{i-j}$$

Skill Measures: There are multiple skill measures (y_{si}) observed for student *i* at the end of ninth grade (such as test scores, grades, etc.). Each scalar skill measure (y_s) is a function of the two-dimensional ability vector (α_{ij}) as in [2], where $\beta_s = (\beta_{cs}, \beta_{ns})^T$ is a vector of "skill prices" describing how each y_s depends on each of the two ability types, and ε_{sij} is a random shock.

[2] $$y_{sij} = \alpha_{ij}{}^T \beta_s + \varepsilon_{sij} \equiv \left(\upsilon_i + \omega_{ij} + \varphi_{i-j}\right)^T \begin{pmatrix} \beta_{cs} \\ \beta_{ns} \end{pmatrix} + \varepsilon_{sij}$$

There is a longer-run outcome (y_l) that policymakers care about (such as high-school graduation or college going), but cannot be measured contemporaneously. The longer-run outcome is also a function of student ability as in [3], where ε_{lij} is random error, and $\beta_{cl} \times \beta_{nl} \neq 0$.

[3] $$y_{lij} = \alpha_{ij}{}^T \beta_l + \varepsilon_{lij} \equiv (\upsilon_i + \omega_{ij} + \varphi_{i-j})^T \begin{pmatrix} \beta_{cl} \\ \beta_{nl} \end{pmatrix} + \varepsilon_{lij} .$$

Teacher Effects**:** Teachers affect student skill measures and outcomes only through their effects on students' accumulated ability. From [2] and [3], teacher *j*'s effect on outcome or skill measure y_z of student *i*, where $z \in \{s, l\}$, is a weighted average of teacher *j*'s effective quality for each dimension of student ability $\theta_{zij} = \omega_{ij}{}^T \beta_z$. Let $\theta_{zj} = E[\omega_{ij}]^T \beta_z$ be the *average* effect of teacher *j* on outcome y_z (i.e. the effect on the average student). Because $E[\omega_{ij}] = E[D_i \omega_j]$, it follows that θ_{zj} is a linear function of the teacher quality vector (ω_j). For expositional purposes, I refer to the teachers' average effect on short-run outcomes (i.e. skill measures) as value-added.[19]

Claim: *If a skill measure reflects a different mix of abilities from that measured by test scores, teachers' value-added on that skill measure may explain variation in teachers' average effects on longer-run outcomes that is not explained by their test-score value-added.*

17 The vector ω_{ij} is a two-dimensional student-specific teacher quality vector. This relaxes the commonly-made assumption that teacher effects are the same for all students (see Jackson et al 2014).
18 Appendix D outlines the explicit production function assumption that justifies the additive model in [1]. I also present empirical evidence to support the assumption of additivity across teachers in Appendix J.
19 This definition is appropriate in the current context because the empirical models employed to estimate teacher effects on ninth-grade skill measures (outlined in Section IV) control for lagged outcomes.

To illustrate this point, consider two ninth-grade skill measures, test scores (y_1) and behaviors (y_2), and a longer-run outcome, high-school graduation (y_l). Assume that value-added on test scores and behaviors are perfect measures (i.e. there is no estimation or measurement error).[20] The best linear unbiased estimate of the average teacher effect on graduation (y_l) based on test-score value-added (θ_{1j}) is $\gamma\theta_{1j}$, where $\gamma = cov(\theta_{lj}, \theta_{1j})/var(\theta_{1j})$. The variation in a teacher's average effect on graduation (θ_{lj}) unexplained by her test-score value-added (θ_{1j}) is a linear function of her quality vector $\ddot{\theta}_{lj} = f(\omega_j)$.[21] Similarly, a teacher's behaviors value-added (θ_{2j}) unexplained by her test-score value-added (θ_{1j}) is a linear function of the *same* teacher quality vector $\ddot{\theta}_{2j} = g(\omega_j)$. Consider the linear regression predicting the average teacher effect on the longer-run outcome (θ_{lj}) as a function of her test-score value-added (θ_{1j}) and her behaviors value-added (θ_{2j}). From Greene (2002), behaviors value-added (θ_{2j}) increase the explained average teacher-level variation in graduation *iff* $cov(\ddot{\theta}_{lj}, \ddot{\theta}_{2j}) \neq 0$.[22] Because both $\ddot{\theta}_{lj}$ and $\ddot{\theta}_{2j}$ are functions of ω_j, it follows that $cov\left(\ddot{\theta}_{lj}, \ddot{\theta}_{2j}\right) \neq 0$, so that behaviors value-added will increase the explained teacher-level variation in graduation.[23] This argument can be applied to any additional skill measure (y_2) and any longer-run outcome (y_l). Note that this result *does not* require that the additional skill measure is unrelated to test scores, but only that there is meaningful variation in abilities measured by the other skill measure that is unrelated to test scores.[24]

---

[20] This assumption is made to highlight the fact that the theoretical result holds *even if* teacher value-added on test scores are perfectly measured.

[21] A teacher's average effect on the long run outcome is $\theta_{lj} = \beta_{cl}\omega_{cj} + \beta_{nl}\omega_{nj}$. The variation in θ_{lj} unexplained by θ_{1j} is $\ddot{\theta}_{lj} = f(\omega_j) = (\beta_{cl} - \gamma\beta_{c1})\omega_{cj} + (\beta_{nl} - \gamma\beta_{n1})\omega_{nj}$. Similarly, the variation in θ_{2j}unexplained by θ_{1j} is $\ddot{\theta}_{2j} = g(\omega_j) = (\beta_{c2} - \pi\beta_{c1})\omega_{cj} + (\beta_{n2} - \pi\beta_{n1})\omega_{nj}$, where $\pi = cov(\theta_{2j}, \theta_{1j})/var(\theta_{1j})$.

[22] See Appendix D for a more formal proof of this statement.

[23] This is also possible if the different teacher effects measure the same skill but are each measured with error. However, in section VI, I demonstrate that this is unlikely to be the case for the outcomes in this paper.

[24] There is an important caveat intrinsic to the use of measurements of behavior as proxies for latent skills. I discuss the short-run outcomes as being pure proxies of skill. However, value-added on the skill-measures may predict impacts on longer-run outcomes through changes in skills but *also* through the effects of the skill-measures directly (a behavior effect). See Heckman (1981 and 1981a) for a discussion of this basic identification problem. For example, consider that dropout is a function of both motivation (an underlying skill) and how far behind a student falls in class (which is a function of absences). A teacher who reduces absences may reduce dropout by (a) increasing student motivation (a skill mechanism) but *also* by (b) reducing the likelihood that a student falls behind (a direct behavior mechanism). To be clear, both mechanisms are causal and each is policy relevant. That is, if teachers systematically increase students' chances of graduating high school in a manner that is not detectable using test-score value-added, irrespective of the mechanism, this would be an important and policy-relevant finding. One implication of this, however, is that teacher value-added on behavior-based skill measures (such as absences and discipline) may better predict teacher impacts on longer-run outcomes than non-behavior-based measures of skill (such as surveys or test scores).

IV Empirical Strategy: Identifying Teacher Impacts on Student Outcomes

This section outlines the model used to estimate teachers' average impacts on student skill measures in ninth grade (θ_{zj}). I refer to these estimated teacher impacts on skill measures as value-added. The value-added estimates are then used as predictors of longer-run outcomes (y_l). From [2], each ninth-grade skill measure y_z for student i with teacher j is a linear function of student ability at the end of ninth grade plus a random error as in [4] below.

[4] $$y_{zij} = (\upsilon_i + \omega_{ij} + \varphi_{i-j})^T \beta_z + \varepsilon_{zij} = {\upsilon_i}^T \beta_z + {\omega_{ij}}^T \beta_z + {\varphi_{i-j}}^T \beta_z + \varepsilon_{zij}.$$

Cross multiplying out terms and substituting in θ_{zj} leads to [5].

[5] $$y_{zij} = \theta_{zj} + \upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz} + \varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz} + \varepsilon_{zij}.$$

When incoming student ability is not observed, and one only observes the value-added of ninth-grade teacher j in a particular subject, [5] becomes [6] below.

[6] $$y_{zij} = \theta_{zj} + u_{zij} \ , \text{ where } u_{zij} = \upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz} + \varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz} + \varepsilon_{zij}.$$

As a normalization, let $E[u_{zij}] = 0$. An estimate of teacher j's value-added on outcome z (θ_{zj}), is the average outcome for all students with teacher j given by $\hat{\theta}_{zj} = \bar{y}_{zi\in j}$. If teachers and students are *both* distributed randomly such that $E[u_{zij}|\theta_{zj}] = E[u_{zij}]\ \forall\, j, \forall\, i$, then, in expectation, the difference in average outcomes for all students with teacher with j and all students with teacher j' will yield the difference in value-added between teacher j and teacher j' for outcome z. That is, $E[\hat{\theta}_{zj} - \hat{\theta}_{zj'}] = \theta_{zj} - \theta_{zj'}$

Because teachers and students are not distributed randomly, differences in teacher-level mean outcomes are unlikely to yield the differences in value-added of individual teachers for two reasons. First, students may sort into schools, and to teachers within schools, by parental socioeconomic status and incoming ability so that $E[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}|\theta_{zj}] \neq E[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}]$. Second, good teachers may cluster in the same schools, and teach the same group of students within schools due to tracking, so that $E[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}|\theta_{zj}] \neq E[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}]$. For example, if good math teachers teach the same group of students as the good English teachers, average classroom outcomes for the math teacher will confound that teacher's value-added with the value-added of the English teacher to which her students are exposed.

To address these two sources of potential bias, I contend that if there exists a set of conditioning variables (T_{ij}) such that (a) students are randomly assigned to teachers, conditional on T_{ij}, and (b) the quality of the teacher of one subject is unrelated to the quality of the teachers of other subjects, conditional on T_{ij}, one can obtain unbiased estimates of the relative value-added of an individual teacher on student outcomes. I outline this logic below.

Identifying assumption 1: *Conditional random assignment of students to teachers*

[7] $$E\left[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}\middle|\theta_{zj}, T_{ij}\right] = E\left[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}\middle|T_{ij}\right] \,\forall\, j, \forall\, z.$$

Conditional on T_{ij}, the value-added of teacher *j* is uninformative about the expected incoming ability of students of teacher *j*.

Identifying assumption 2: *Conditional independence of teacher effects*

[8] $$E\left[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}\middle|\theta_{zj}, T_{ij}\right] = E\left[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}\middle|T_{ij}\right] \,\forall\, j, \forall\, z.$$

Conditional on T_{ij}, the value-added of teacher *j* is uninformative about the value-added of *other* teachers (of different subjects) of the students of teacher *j*.

Even though $E\left[\hat{\theta}_{zj}\right] = \theta_{zj} + E\left[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}\middle|T_{ij}\right] + E\left[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}\middle|T_{ij}\right] \neq \theta_{zj}$, under assumptions 1 and 2, $E\left[\hat{\theta}_{zj} - \hat{\theta}_{zj'}|T_{ij}\right] = \theta_{zj} - \theta_{zj'}$. That is, for a given outcome z, even with sorting of students to teachers and clustering of teachers to groups of students, in expectation, the difference in mean outcomes for teacher *j* and that for teacher *j'* conditional on T_{ij} will yield the difference in value-added between teacher *j* and teacher *j'*.

The proposed T_{ij} includes several variables. To account for student ability sorting, I include two lags of math scores, English scores, repeater status, suspensions, and attendance, and a single lag of GPA.[25] To account for sorting that occurs at the group level, I include classroom averages of the eighth-grade skill measures and demographics (Protic et al. 2013). To account for sorting of teachers to groups of classes such that teacher quality may be correlated across subjects for the same student, I control for the number of honors courses taken (Harris and Anderson 2012; Aaronson, Barrow, and Sander 2007), and I include fixed effects for the student's school track (Jackson 2014). The school track is the unique combination of the ten core academic courses, the level of math taken, and the level of English taken *in a particular school*.[26] Only students at the same school who also take the same academic courses, level of English, and level of math are in the same school track.[27] I refer to the school track as "track" for the remainder of the paper.

[25] Kane and Staiger (2008) and Kane et al. (2013), find that inclusion of one year of lagged outcomes is sufficient to eliminate bias due to sorting. Rothstein (2010) advocates using two lags.

[26] Defining tracks flexibly at the school/course-group/course level allows for different schools that have different selection models and treatments for each track. See Appendix E for further discussion of tracks.

[27] Students taking the same courses at different schools are in different school-tracks. Students at the same school in at least one different academic course are in different school tracks. Similarly, students at the same school taking the

The idea behind conditioning on track is as follows: If better teachers sort into particular tracks, the advanced track for example, then students in the advanced track will have both better English and math teachers than those in the regular track. This makes it difficult to disentangle the value-added of the English teachers from that of the math teachers when making comparisons across tracks. However, if teachers sort into tracks but *do not* sort into classes within tracks, then within a track, students with better English teachers are not systematically exposed to better math teachers and vice versa. This allows one to isolate the value-added of one subject teacher from that of teachers in other subjects within tracks. Importantly, if students sort into tracks, making comparisons among students within tracks will also help eliminate student sorting bias.

In sum, if students are randomly assigned to classrooms within tracks (conditional on the rich set of controls), then conditional on tracks and controls, Identifying Assumption 1 will be satisfied. Similarly, if teachers are randomly assigned to classrooms within tracks (conditional on the rich set of controls), then conditional on tracks and controls, Identifying Assumption 2 will be satisfied. I present evidence to support the validity of these identifying assumptions in section VI.

IV.A Identifying Teacher Impacts on Ninth-Grade Skill Measures

I follow the convention in the teacher value-added literature and model outcome (or skill measure) *z* of student *i* in classroom *c* with teacher *j* in school *s* in year *t* with equation [9].

[9] $$y_{zicjst} = \Omega_z X_{icjst} + \tau_{st} + e_{zicjst}$$

Here, X_{icjst} includes all the time-varying variables in T_{ij} discussed above, and τ_{st} are school-by-year indicator variables to account for transitory school-level shocks. Removing the influence of observables yields $e_{zicjst} = y_{zicjst} - \Omega_z X_{icjst} - \tau_{st}$. This student-level residual is comprised of teacher value-added (θ_{zj}), a random classroom-level shock (ε_{zcjst}), and random student-level error (ε_{zicjst}), such that $e_{zicjst} = \theta_{zj} + \varepsilon_{zcjst} + \varepsilon_{zicjst}$. The average of these student-level residuals over time for a given teacher *j* is connoted $\bar{e}_{zj}$, and is an unbiased estimate of teacher *j*'s value-added on outcome *z* under the aforementioned identifying assumptions.

Even though $\bar{e}_{zj}$ is an unbiased estimate of teacher *j*'s value-added on outcome *z*, to avoid mechanical endogeneity, one should not estimate teacher value-added using the same students among whom longer-run outcomes are being compared. Accordingly, I follow Chetty, Friedman,

same courses but taking the same math or English class at different levels are in different school tracks. Because many students pursue the same course of study, less than 1 percent of all students are in singleton tracks, 83 percent of students are in tracks with more than 20 students, and the median student is in a school track with 199 other students.

and Rockoff (2014a) and *predict* how much each teacher improves student outcomes in a given year based on her performance in *other* years (with a different set of students). This leave-year-out (jackknife) measure of teacher quality removes the endogeneity associated with using the same students to form both the treatment and the outcome, and isolates the variability in teacher value-added that persists over time. A leave-year-out estimate for teacher *j* in year *t* is the teacher's average residual based on all other years of data (-*t*) as below (equation [10]).

[10] $$\hat{\theta}_{zj,-t} = \bar{e}_{zj,-t}.$$

The estimate, $\hat{\theta}_{zj,-t}$, minimizes mean square *estimation* error and is an unbiased estimate of θ_{zj}. However, because $\hat{\theta}_{zj,-t}$ is estimated with noise, $\hat{\theta}_{zj,-t}$ is not the optimal out of sample predictor and does not minimize out-of-sample *prediction* errors. To minimize mean squared prediction errors, it is optimal to introduce some bias and to use the raw estimates to form empirical Bayes (or shrinkage) estimates (Kane and Staiger 2008; Chetty, Friedman, and Rockoff 2014a; Gordon, Kane, and Staiger 2006).[28] This approach models the estimation error in each teacher's raw mean and shrinks noisier estimates towards the grand mean (in this case, zero). The resulting leave-year-out empirical Bayes estimate of teacher *j's* value-added is described by [11].

[11] $$\hat{\mu}_{zjt} = \hat{\theta}_{zj,-t}\lambda_{zj}.$$

This empirical Bayes estimate for each teacher's value-added is the leave-year-out teacher-level mean ($\hat{\theta}_{zj,-t}$) multiplied by λ_{zj}, an estimate of its reliability.[29] As a result, less reliable

[28] Though these are commonly referred to as estimates, they are really predictors. The best linear *predictor* of student outcomes given the leave-year-out teacher effect is obtained from a regression of y on $\hat{\theta}_{zj,-t}$. That is $\mathrm{E}(y_{zicjst}|\hat{\theta}_{zj,-t}) = a + b(\hat{\theta}_{zj,-t})$. Where the estimates effects are normalized to be mean 0, it follows that a=0. Because the effects are estimated with error, it follows that $b = \mathrm{var}(\theta_{zj,})/\mathrm{var}(\hat{\theta}_{zj,-t}) < 1$. Even though $\hat{\theta}_{zj,-t}$ is an unbiased estimate of θ_{zj}, the optimal predictor that minimizes prediction errors is $b(\hat{\theta}_{zj,-t})$.

[29] Following Kane and Staiger (2008), Gordon, Kane, and Staiger (2006), Jackson (2013), and Jackson and Bruegmann (2009), $\lambda_{zj} = \left[\frac{\sigma^2_{\theta_{zj}}}{\sigma^2_{\theta_{zj}} + \left(\sum_{m_j}(1/(\sigma^2_{\varepsilon_{zcjst}} + \sigma^2_{\varepsilon_{zicjst}}/n_{cj})\right)^{-1}}\right]$ where n_{cj} is the number of students in class *c* with teacher *j*, and m_j is the number of classrooms for teacher *j*. The parameters $\sigma^2_{\theta_{zj}}$, $\sigma^2_{\varepsilon_{zcjst}}$, and $\sigma^2_{\varepsilon_{zicjst}}$ are replaced by empirical estimates under the assumption $cov(\theta_{zj}, \varepsilon_{zcjst}) = cov(\theta_{zj}, \varepsilon_{zicjst}) = cov(\varepsilon_{zicjst}, \varepsilon_{zcjst}) = 0$. Under this assumption, $var(e_{zicjst}) = \sigma^2_{\varepsilon_{zicjst}} + \sigma^2_{\varepsilon_{zcjst}} + \sigma^2_{\theta zj}$ and $cov(\bar{e}_{zcjt}, \bar{e}_{zc'jt'}) = \sigma^2_{\theta zj}$ where $\bar{e}_{zcjt}$ is the average residual for classroom *c* for teacher *j* in year *t* and $\bar{e}_{zc'jt'}$ is the average residual for classroom *c'* for teacher *j* not in year *t*. As such, $\sigma^2_{\varepsilon_{zicjst}}$, the empirical estimate of the variance of the student-level errors, is estimated using the sample variance of the student-level residuals within classrooms. Also $\sigma^2_{\theta_{zj}}$, the empirical estimate of the variance of the true teacher value-added on outcome *z*, is estimated using the sample covariance of classroom-level mean residuals for the same teacher in *different* years. Under the assumptions above, I can obtain an empirical estimate of $\sigma^2_{\varepsilon_{zcjst}}$, the variance of

estimates (i.e. those that are estimated with more noise due to a small number of students, or a small number of classrooms, or both) are shrunk toward the grand mean for all teachers.[30] To examine whether teacher value-added on test scores and behaviors predict teacher impacts on longer-run outcomes, I use the estimates from [11] as predictors of the longer-run outcomes. In all the empirical sections of this paper, when I refer to value-added estimates, I am referring to the leave-year-out empirical Bayes estimates as in [11].

V Effects on Skill Measures

Before presenting impacts on longer-run outcomes, I examine the magnitudes of teacher value-added on the proposed skill measures (i.e. test scores and behaviors). I follow Kane and Staiger (2008) and for each outcome use the covariance between mean classroom residuals for the same teacher as a measure of the variance of the persistent component of teacher value-added ($\hat{\sigma}^2_{\theta_{zj}}$).[31] The square root of estimated variances (i.e. the implied standard deviations of teacher value-added) for all ninth-grade outcomes are presented for each subject in table 3.

The standard deviation of the math teacher value-added on math test scores is 0.084σ so that having a math teacher with value-added at the 85th versus 50th percentile on math test scores would increase math scores by roughly 0.084σ. The relationship between average test scores and graduation in column 4 of table 2 implies that this would be associated with a 0.16 percentage point increase in the likelihood of high-school graduation. Looking to value-added on behaviors, having a math teacher at the 85th versus 50th percentile reduces the likelihood of being suspended by 1.2 percentage points, has no impact on absences, increases GPA by 0.063 grade points, and increases on-time grade progression by 2.64 percentage points. Combining the ninth grade

---

the classroom-level shocks, using the variance of the total residual, $var(e_{zicjst})$, minus the empirical estimates of $\sigma^2_{\varepsilon_{zijst}}$ and $\sigma^2_{\theta_{zj}}$.

30 Teachers with no estimated raw fixed effects (i.e. those in the data for only one year) are shrunk toward the mean of other teachers with similar observable attributes. Teachers with missing estimates are given the fitted value from a regression predicting $\hat{\mu}_{zjt}$ based on observable teacher characteristics (gender, ethnicity, experience, certification, license status, college selectivity, and test scores). Teachers for whom there are no observable characteristics are given the mean of the distribution of the estimated $\hat{\mu}_{zjt}$. Results are very similar to those obtained when the teacher estimates are shrunk to zero for teachers with no estimated out of sample effect.

31 Under the identifying assumptions, $cov(\bar{e}_{zcjt,}, \bar{e}_{zc'jt'}) = \sigma^2_{\theta zj}$ where $\bar{e}_{zcjt}$ is the average residual for classroom c for teacher j in year t and $\bar{e}_{zc'jt'}$ is the average residual for classroom c' for teacher j not in year t. To estimate $\sigma^2_{\theta zj}$, I compute mean residuals ($\bar{e}_{zcjt}$) for each classroom. Then I pair every classroom with another random classroom for the same teacher ($\bar{e}_{zc'jt'}$) and compute the covariance of the mean residuals across these classrooms. I replicate this procedure 200 times and take the median of the estimated covariance as the parameter estimate.

behaviors into a single variable, having a math teacher at the 85th versus 50th percentile of value-added on the behavior index would increase the behavior index by 0.08σ. The relationships in table 2 suggest that this would lead to a 1.27 percentage point increase in the likelihood of high-school graduation. Patterns for English teachers are similar. However, as in other settings, value-added on English scores are smaller than those on math scores (see Jackson, Rockoff, and Staiger 2014). The correlations indicate that having an English teacher with value-added at the 85th versus 50th percentile on English scores would increase English scores by 0.03σ. Having an English teacher with value-added at the 85th versus 50th percentile on the behavior index would increase the behavior index by roughly 0.055σ— an effect size on behaviors that is on the same order of magnitude as those for math teachers. The patterns presented in table 3 indicate that there is economically meaningful variation in outcomes across teachers that persists across classrooms.

One may worry that these correlations are driven by systematic reporting bias (e.g. teachers who are easy graders or do not report students to the principal's office may mechanically appear to improve student outcomes without actually improving underlying behaviors). Because passing English and math is required to graduate from high school, and an expelled student will not graduate, such reporting biases could mechanically improve graduation and reduce dropout without any real skill improvement or improvement in behaviors. However, ninth grade teachers who systematically raise students' course grades in tenth grade (when they are no longer directly interacting with the student) cannot be doing so by being easy graders or by being more likely to punish students. If ninth grade teachers who systematically improve GPA grades in 10th grade also improve longer-run outcomes, it will likely be through improvements in student skill (rather than any mechanical grade inflation effects or reporting biases). As a robustness check, to provide a measure of teacher value-added on noncognitive skills that is not subject to grading or reporting biases, I also present results using ninth-grade teacher value-added on tenth-grade GPA as a proxy for teacher effects on noncognitive skills (last column). For both math and English teachers there is systematic ninth-grade teacher-level variation in tenth-grade GPA. The implied standard deviation of teacher value-added is 0.05 grade points for math and 0.026 for English. This indicates that ninth-grade teachers impact behavior-based measures of noncognitive skills that are not due to reporting or grading standards. Whether this teacher-level variation can be well-measured for individual teachers, and whether estimated teacher value-added on the different skill measures reflect effects on different skills are explored below.

V.A Relationship between Teacher Effects Across Skill Measures

To explore whether teachers who improve test scores improve other skill measures, table 4 presents the raw correlations between the value-added estimates on the different skill measures, where the data for both math and English teachers are combined. Teachers with higher test-score value-added are associated with better non-test-score outcomes, but the relationships are weak. The correlations between test-score value-added and that on being suspended or absences are both below 0.1. The test-score value-added estimates are somewhat more highly correlated with value-added on GPA (*r*=0.22) and on-time grade progression (*r*=0.16), but not strongly so. The correlation between test-score value-added and that on the behavior index is only 0.15 such that less than 3 percent of the variation in teacher value-added on the behavior index is associated with teacher value-added on test scores, and vice versa. Looking to impacts on tenth-grade GPA (which is free from reporting and grading biases), the correlation with test-score value-added is only 0.11. However, because the value-added estimates are estimated with noise, the variation in value-added on behaviors that is unrelated to value-added on test-scores may simply reflect statistical noise, and not systematic variation in teacher quality *per se*.

To further explore whether teacher's behaviors value-added reflect impacts on skills that are unmeasured by their test-score value-added, I regress student skill measures on their teachers' leave-year-out value-added estimates for those skill measures. Specifically, I estimate equation [12] below where all variables are defined as in [9] and $\hat{\mu}_{test,jt}$ and $\hat{\mu}_{behavior,jt}$ are the leave-year-out empirical Bayes value-added estimates on test scores and the behaviors, respectively.

[12] $$y_{zicjst} = \Omega_z X_{icjst} + \delta_{z1} \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \delta_{z2} \cdot (\varrho_2 \hat{\mu}_{behavior,jt}) + \delta_d \sum_{d=1}^{4} I_d + \nu_{zicjst}.$$

For ease of interpretation, the teacher value-added estimates are multiplied by scaling factors ϱ_1 and ϱ_2 so that the coefficients δ_1 and δ_2 identify the effect of increasing teacher value-added on test scores and the behaviors, respectively, by one standard deviation (as presented in table 3).[32] Data for all subjects are stacked and the results are presented for both subjects combined. All models include indicators for the specific course of the teacher (I_d) (i.e. English, geometry, algebra I, and algebra II). To account for the fact that individual students enter the stacked dataset in both

[32] To obtain the scaling index for each outcome I first estimate equation [a] below for each outcome z.

[a] $$y_{zicjst} = \beta_z X_{icjst} + \pi_z \cdot \hat{\mu}_{zjt} + \nu_{zicjst}$$

The scaling index is $\varrho_z = |\hat{\pi}_z / \hat{\sigma}_{\theta_{zj}}|$, where $\hat{\pi}_z$ is the coefficient estimate from [a] and $\hat{\sigma}_{\theta_{zj}}$ is the estimated standard deviation of true teacher value-added on outcome z described in table 3. This rescaling is done separately by subject.

subjects and individual teachers have multiple students, standard errors are adjusted for two-way clustering at the teacher and student levels following Cameron, Gelbach, and Miller (2011).

Table 5 presents the coefficients on the rescaled value-added estimates. As expected, columns 1, 7, and 10 show that teachers who raise a given skill measure *out of sample* have large statistically significant effects on that same skill measure. Increasing teacher test-score value-added (across both subjects) by one standard deviation increases test scores by 0.0685σ (*p*-value<0.01),[33] increasing teacher behaviors value-added by one standard deviation increases the behavior index by 0.0579σ (*p*-value<0.01), and increasing teacher tenth-grade GPA value-added by one standard deviation increases tenth-grade GPA by 0.0357σ (*p*-value<0.05). Consistent with the teacher value-added on the skill measures being positively correlated, columns 2, 6, and 9 reveal that teachers who raise test scores improve behaviors and vice versa.

However, models that include value-added on both kinds of skill measures simultaneously suggest that teacher value-added on behaviors capture impacts on skills that are unmeasured by tests. Specifically, conditional on teachers' test score value-added, behaviors value-added is strongly predictive of improved behaviors (column 8), but weakly associated with *lower* test scores (column 3). Similarly, conditional on teachers' test score value-added, teacher value-added on 10th grade GPA is strongly predictive of tenth-grade GPA (column 11), but unrelated to test scores (column 5). That is, conditional on teacher test-score value-added, teacher value-added on behaviors predict large improvement on behaviors but no improvement in test scores. As such, a teacher's behaviors value-added likely captures effects on noncognitive skills that are not well-measured by test scores. If so, as indicated by the model, behaviors value-added may explain teachers' impacts on longer-run outcomes that are not measured by their test score value-added.

VI Predicting Longer-Run Teacher Impacts with Value-Added on Skill Measures

This section tests whether teachers who improve behaviors *cause* improved longer-run outcomes (conditional on their test-score value-added). To this aim, I estimate equation [12] in which the outcomes are measures of high-school completion; whether the student subsequently dropped out of secondary school by twelfth grade, and whether they graduated from high school. For ease of interpretation, I present estimates of the average marginal effects using linear probability models. Because linear models can be misleading about marginal effects for binary

---

[33] The table presents the effects for math and English test scores combined. As such, the pooled effect across both subjects lies between the estimated standard deviation for math teachers (0.084), and that for English teachers (0.03).

outcomes, I also present conditional logit estimates and the ensuing implied marginal effects. To quantify the increase in the ability to predict variability in teachers' longer-run impacts by adding behaviors value-added, using the linear model, I estimate [12] both with and without behaviors value-added, and I compute the percentage increase in the predicted variance of the teacher effects on the longer-run outcomes.[34] The results are presented in table 6.

Column 1 presents the effect of increasing test-score value-added on high-school graduation when behaviors value-added is not included. On average, one standard deviation higher test-score value-added leads to a 0.152 percentage point increase in high-school graduation (p-value<0.01). To put this estimated effect into perspective, the linear relationship between a one standard deviation increase in test scores and graduation in table 2 (1.86 percentage points) multiplied by the estimated standard deviation of test-score value-added in table 3 (0.075) implies that a one standard deviation increase in teacher test-score value-added would increase high-school graduation by 1.86*0.075=0.139 percentage points. This is very close to the estimated magnitudes—suggesting that the results are reasonable.

Column 2 presents the teacher effects on high-school graduation using both teacher value-added on the behavior index and test scores. Given that the two are weakly correlated, the point estimate for test-score value-added remains largely unchanged. Conditional on a teacher's behaviors value-added, increasing test-score value-added by one standard deviation increases high-school graduation by 0.118 percentage points, and conditional on a teacher's test-score value-added, increasing a teacher's behaviors value-added by one standard deviation increases high-school graduation by 1.46 percentage points (p-value<0.01). The linear relationship between a one standard deviation increase in behaviors and graduation in table 2 (15.8 percentage points) multiplied by the estimated standard deviation of teacher behaviors value-added in table 3 (0.0769) implies that a one standard deviation increase in a teacher's behaviors value-added would increase students' high-school graduation by 15.8∗0.0769=1.215 percentage points. This is very close to the estimated magnitudes, —suggesting that the results are reasonable and that the magnitudes are

[34]Specifically, I estimate both [b] and [c] below

[b] $y_{zicjst} = \Omega_z X_{it} + \delta_{z1} \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \delta_d \sum_{d=1}^{4} I_d + \nu_{zicjst}$.

[c] $y_{zicjst} = \Omega_z X_{it} + \delta_{z1} \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \delta_{z2} \cdot (\varrho_2 \hat{\mu}_{behavior,jt}) + \delta_d \sum_{d=1}^{4} I_d + \nu_{zicjst}$.

I compute the variance of the fitted values for each teacher from both models. In models without behaviors value-added (i.e. [b]) this is $a = var(\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}))$, and in models with teacher value-added on both (i.e. [c]), this is $b = var[\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \hat{\delta}_2 \cdot (\varrho_2 \hat{\mu}_{behavior,jt})]$. The percentage increase in explained variance from also including behaviors value-added (versus using their test-score value-added alone) is $100 \times ((b \div a) - 1)$.

plausible. Comparing the estimated teacher-level variability in high-school graduation from the fitted models with both value-added estimates to those using only test-score value-added, including behaviors value-added increases the explained variance of teacher effects on graduation by 305% percent – i.e. more than quadruples the variance of the identifiable teacher effect on high-school graduation. This is consistent with Chamberlain (2013) who finds that test-score value-added may account for less than one fifth of the overall effect of teachers on college-going.

Column 4 presents the results from a conditional logit specification to more accurately reflect the binary outcome. The estimated coefficient estimates are presented, with standard errors below in brackets, and the average marginal effects are presented in parenthesis below that.[35] In models with teacher value-added on both test scores and behaviors, increasing test-score value-added by one standard deviation increases high-school graduation by 0.2 percentage points (*p*-value>0.1), and increasing the teacher's behaviors value-added by one standard deviation increases high-school graduation by 3.31 percentage points (*p*-value<0.01). Even though the linear and nonlinear models yield somewhat different marginal effects, they are on the same order of magnitude and have overlapping 95% confidence intervals. To put these effect sizes into perspective, consider the following back-of-the-envelope calculation. In the linear model, increasing a teacher's behaviors value-added by one standard deviation increases high-school graduation by 1.46 percentage points, on average. The average teacher has 54.5 students a year. According to the Bureau of Labor Statistics (United States Department of Labor 2016), completing high school is associated with $11,000 higher annual earnings. Assuming this difference is causal, increasing high-school graduation rates by 1.46 percentage points would increase annual earnings by roughly $160 per year per student. This figure multiplied by 54 students is $8,670 higher cohort earnings each year. Assuming this increase stays the same each year for forty years, at a 7% discount rate, this translates into $126,286 in present discounted lifetime earnings per year of students taught. Even though some of the raw differences in earnings assumed in this rough calculation may reflect selection, under most reasonable assumptions regarding the economic benefits of completing high school, the estimated effects are economically important.

The other measure of school completion is high-school dropout. High-school dropout is notoriously difficult to measure (Tyler and Lofstrom 2009) so that the effects will likely be muted.

[35] Because the conditional logit model conditions on track, marginal effects cannot be estimated directly. The reported marginal effects are approximate and are computed assuming that the track effects are equal to zero.

However, it is helpful to show that the same patterns hold for both high-school graduation and high-school dropout. Column 7 shows that, on average, a one standard deviation increase in teacher test-score value-added reduces the likelihood of dropping out by 0.03 percentage points, and a one standard deviation increase in teacher behaviors value-added reduces the likelihood of dropout by 0.4 percentage points (p-value<0.05). While the point estimates from the linear model are smaller than those for graduation, the implied marginal effects from the conditional logit model are similar across the two outcomes. In models with value-added on both test scores and behaviors (column 9), a one standard deviation increase in teacher value-added on test scores and behaviors reduces the likelihood of dropout by 0.12 and 2.71 percentage points, respectively. As with high-school graduation, including teachers' behaviors value-added increases the explained teacher-level variance in dropout by 326% percent. The consistency across both measures of school completion suggests that the estimated effects reflect real changes in human capital acquisition. Note that if teachers impact skills not captured by test scores or the behaviors (which is likely), the estimates presented may still understate teacher's full effect on longer-run outcomes.[36]

One may worry that the results presented thus far could emerge even if there were no improvement in skills or improvement in behaviors if some teachers are easy graders or less likely to report student's poor behavior. To assuage such concerns, I present the same set of results using the ninth-grade teacher's value-added on *tenth-grade* GPA instead of the ninth-grade behaviors (columns 3, 5, 8 and 10). While the standard errors are larger, the parameter estimates are almost identical to those using value-added on ninth-grade behaviors. In the linear models that include teacher test-score value-added, a one standard deviation increase in teacher value-added on 10^{th} grade GPA is associated with a 1.46 percentage point increase in high-school graduation and a 0.3 percentage point reeducation in high-school dropout. Ninth grade teachers' reporting or grading biases do not influence students 10^{th} grade GPA. As such, the results presented are not mechanical or driven by reporting bias, and reflect teachers either inducing real improvement in student skills or promoting productive behaviors that improve students' longer-run outcomes.[37]

[36] As mentioned in Section III, these causal effects may reflect a pure skill effect or a behavioral effect due to improved behaviors themselves. If the behavioral effects (such as being in class more often directly causing students to stay in school) are larger for the noncognitive proxies than for test scores, it could partially explain why behaviors value-added predict larger impacts on dropout and graduation than test-score value-added. Irrespective of the mechanism, the effects are causal and the larger impact of behaviors value-added is policy relevant.

[37] Appendix F present results using teacher value-added on each behavior individually. Teacher value-added on individual behaviors have the expected signs and many are statistically significant. Because eighth-grade GPA is imperfectly measured, I show that the results are robust to using teacher value-added on an index that excludes GPA

VI.A Testing the Identifying Assumptions

The first identifying assumption is that students are randomly assigned to teachers conditional on observables.[38] To present evidence that this condition is satisfied, I first implement a test for selection on observables (Appendix G). I show that conditional on eighth-grade outcomes and controls for tracks, teacher value-added estimates are unrelated to their students' predicted dropout and predicted graduation (weighted indices of parental education, gender, ethnicity, and seventh-grade math scores, reading scores, grade repetition, suspensions, and absences). To test for selection on unobservables within school cohorts, I follow Chetty, Friedman, and Rockoff (2014b) and exploit the statistical fact that the effects of any selection among students within a cohort at a given school will be eliminated by aggregating the treatment to the school-year level and relying only on cohort-level variation across years within schools. That is, if value-added estimates merely capture selection within school cohorts, then the arrival of a teacher who increases the average teacher estimated value-added for a cohort but has no effect on real teacher quality should have no effect on average student outcomes for that cohort. Conversely, if the value-added estimates reflect real impacts, differences in average estimated teacher value-added *across* cohorts (driven by changes in teaching personnel within schools over time) should be associated with similar outcome differences as similar differences in estimated value-added across individual students *within* cohorts. To test for this, I implement instrumental variables models that use only variation across cohorts within a school (Appendix G). The main findings are robust to using the clean variation *across* cohorts. In sum, I find no evidence of selection bias so that the first identifying assumption is likely valid.

The second identifying assumption is that, conditional on observables, the quality of a student's teacher in one subject is unrelated to the quality that student's teachers in other subjects. I test this assumption in two ways (Appendix G). First, for each student I correlate the estimated math and English teacher value-added. The correlation between the math and English teacher test-score value-added is 0.008, that for math and English teacher behaviors value-added is 0.0078, and that for math and English teacher effects tenth-grade GPA value-added is 0.0087. In a

as a skill measure entirely. In sum, Appendix F shows that the value-added on no single behavior drives the effects, and that it is the shared variability across the behaviors (which I posit is due to non-cognitive skills).

[38] Rothstein (2010) argues that teacher value-added models may be biased because students within a cohort within a school may select (or be assigned) to teachers on dimensions that are unobserved by researchers. However, Kane and Staiger (2008), Kane et al. (2013), Chetty, Friedman, and Rockoff (2014b), and Bacher-Hicks, Kane, and Staiger (2015) show that teacher value-added exhibits no appreciable bias in experimental and quasi-experimental data.

regression predicting the math teacher's value-added as a function of the English teacher's value-added, all coefficients are close to zero and all have *p*-values larger than 0.1. I also test whether the main results are robust to the inclusion of fixed effects for the other subject teachers and school-by year fixed effects (to account for subject teachers other than math and English).[39] Linear probability models that include other subject teacher fixed effects and school-by year fixed effects are almost identical to those that do not. This is consistent with no conditional correlation between the quality of teachers across subjects, suggests that the empirical strategy isolates the contribution of the individual teachers, and suggests that the second identifying assumption is valid.

VI.B Effects on Other Outcomes and Predictors of Longer-Run Success

While high-school dropout and graduation are the main longer-run outcomes in this study, I present effects of ninth-grade teachers on a few intermediate outcomes and measures of college going (table 7). I focus attention on the impacts of teachers' behavior value-added conditional on test-score value-added. Consistent with the graduation and dropout results, conditional on teachers' test-score value-added, a one standard deviation increase in behaviors value-added increases enrolling in tenth grade by 2 percentage points (p-value<0.1), increases tenth-grade GPA by 0.013 grade points (p-value<0.1), increases SAT-taking by 1.16 percentage points (p-value<0.1), increases the likelihood of reporting plans to attend a four-year college after high-school graduation by 1.03 percentage points (p-value<0.05), and increases graduating high-school GPA by 0.0214 points (p-value<0.01). The one outcome for which behaviors value-added adds no explanatory power is total SAT score (which *is* affected by a teacher's test-score value-added).

Similar to the patterns for high-school completion, including teachers' behaviors value-added increases the identifiable teacher-level variance by 793 percent for tenth-grade enrollment, 33 percent for tenth-grade GPA, 305 percent for graduation, 326 percent for dropout, 228 percent for SAT-taking, 193 percent for four-year college intentions, and 607 percent for high-school GPA. The lower panel presents results using value-added on 10^{th} grade GPA to assuage concerns regarding reporting biases and mechanical effects. The results are less precise, but similar to those found using ninth-grade behaviors value-added. In sum, teachers' behaviors value-added improve the ability to identify teachers who improve a variety of longer-run outcomes considerably.[40]

---

[39] I augment Equation [12] to include indicator variables for each math (or English) teacher when predicting the impact of the English (or math) teachers on longer-run outcomes.

[40] An exploration of differences by subject is presented in Appendix H. Overall one cannot reject the null hypothesis of no differences across subjects at traditional levels of significance.

VI.3 Possible Policy Uses of Effects on Behaviors

I briefly discuss potential applications of teacher behaviors value-added to policymaking. One possibility would be to identify observable teacher characteristics associated with behaviors value-added and select teachers with these characteristics. To determine the scope of this type of policy, I regress the behavior index on observable teacher characteristics while controlling for school tracks, year effects, and student covariates (table I1). While observable teacher characteristics predict effects on test scores, none of the observable teacher characteristics — years of teaching experience, full certification, teaching exams scores, regular licensing, and college selectivity (as measured by the 75th percentile of the SAT scores at the teacher's college) —are significantly related to behaviors.[41] However, this does not preclude the use of more detailed teacher information to better predict teacher effects on a broad range of skills.

Another policy application is to provide incentives for teachers to improve behaviors. However, because some of the behaviors can be "improved" by changes in teacher behavior that do not improve student skills or behaviors (such as inflating grades and misreporting misconduct) attaching external stakes to the behavior index may not improve student skills. There are three feasible solutions to this "gameability" problem. One possibility is to find measures of noncognitive skills that are difficult to adjust unethically. For example, classroom observations and student and parent surveys may provide valuable information about student skills not measured by test scores and are less easily manipulated by teachers. One could attach external incentives to both these measures of noncognitive skills and test scores to promote better longer-run outcomes. Another approach is to provide teachers with incentives to improve the behaviors of students in their classrooms the *following* year, when the teacher's influence may still be present, but the teacher can no longer manipulate student behaviors (as in Carrell and West 2010 and Figlio, Schapiro, and Soter 2015). A final solution is to identify teaching practices that improve behaviors and provide incentives for teachers to engage in these practices. Such approaches have been used successfully to increase test scores (Taylor and Tyler, 2012; Allen et al. 2011). In sum, the teacher effects on the behaviors used in this study can be useful for policy.

[41] The lack of an experience gradient may seem surprising. However, test-based accountability creates incentives to improve test scores but not behaviors. As such, one might expect an experience gradient for test scores but not for the behavior index. In fact, if teachers can improve test scores by expending less effort on improving behaviors, one might observe a positive experience gradient for test scores and a *negative* one for behaviors.

VII Conclusions

This paper extends the traditional test-score value-added model of teacher quality to allow for the possibility that teachers affect a variety of student outcomes through their effects on both students' cognitive and noncognitive skills. In this model, teachers may have effects on skills that affect longer-run outcomes, are not reflected in test scores, but are reflected in *other* skill measures. I use an index of behaviors to proxy for noncognitive skills and find that ninth-grade teachers have meaningful impacts (i.e. value-added) on both test scores and these behaviors. While test scores and behaviors are positively correlated, value-added on behaviors explain significant variability in teacher impacts on high-school graduation and dropout that are not captured by their test-score value-added. Adding teachers' behaviors value-added more than doubles the identifiable teacher-level variability on longer-run outcomes such as high-school graduation, SAT-taking, and intentions to attend college.

Importantly, to ensure that these patterns reflect real improvement in overall skills, rather than simply reflecting mechanical effects due to grade inflation or reporting bias, I document that teachers who improve behaviors also improve longer-run outcomes that have no mechanical relationship with the behaviors such as SAT-taking or tenth-grade GPA. Moreover, to rule out any mechanical effects, I show that I can replicate all the main patterns using ninth-grade teachers' value-added on tenth-grade GPA (for which there should be no mechanical bias due to reporting or grade inflation). I also present several tests indicating that the effects are real. Overall, the results highlight the fact that using non-test-score skill measures (i.e., behavior measures) to proxy for important noncognitive skills can be fruitful in evaluating teachers specifically and human capital interventions more broadly.

The results provide hard evidence that teacher effects on test scores capture only a fraction of their impact on human capital. Further work is needed to derive measures of those important skills that are not well-captured by standardized tests and difficult for teachers to manipulate. The patterns presented suggest that the resulting gains in student skills and overall well-being may be considerable.

TABLE 1

SUMMARY STATISTICS OF STUDENT DATA

Variable	Obs.	Mean	S.D.	S.D. within Schools	S.D. within Tracks
Math z-score 8th grade [a]	573963	0.233	(0.940)	(0.853)	(0.605)
Reading z-score 8th grade [a]	573963	0.217	(0.943)	(0.882)	(0.678)
Repeat 8th grade	570850	0.006	(0.080)	(0.079)	(0.073)
Suspended (8th Grade)	573963	0.039	(0.193)	(0.191)	(0.180)
Absences (8th Grade)	573963	4.593	(5.655)	(5.593)	(5.196)
GPA (8th Grade) [b]	148464	3.204	(0.846)	(0.515)	(0.374)
Student: Female	573963	0.504	(0.500)	(0.499)	(0.479)
Student: Black	573963	0.261	(0.439)	(0.403)	(0.362)
Student: Hispanic	573963	0.072	(0.259)	(0.255)	(0.241)
Student: White	573963	0.588	(0.492)	(0.446)	(0.402)
Student: Asian	573963	0.021	(0.142)	(0.140)	(0.133)
Parental education: Some High School	573963	0.066	(0.249)	(0.246)	(0.233)
Parental education: High School Grad	573963	0.394	(0.489)	(0.476)	(0.448)
Parental education: Trade School Grad	573963	0.016	(0.126)	(0.126)	(0.122)
Parental education: Community College Grad	573963	0.133	(0.340)	(0.338)	(0.327)
Parental education: Four-year College Grad	573963	0.227	(0.419)	(0.410)	(0.386)
Parental education: Graduate School Grad	573963	0.067	(0.251)	(0.244)	(0.230)
Number of Honors classes	573963	1.545	(1.814)	(1.602)	(0.649)
Algebra I z-Score (9th grade) [a]	358315	0.198	(0.967)	(0.898)	(0.768)
English I z-Score (9th grade) [a]	569705	0.203	(0.922)	(0.858)	(0.645)
Geometry z-Score (9th Grade) [a]	113693	0.061	(0.968)	(0.832)	(0.670)
Algebra II z-Score (9th Grade) [a]	34927	0.087	(0.956)	(0.785)	(0.642)
Math z-score (9th grade)	477524	0.183	(0.959)	(0.900)	(0.751)
Absences (9th Grade)	573963	3.462	(4.991)	(4.902)	(4.430)
Suspended (9th Grade)	573963	0.051	(0.220)	(0.217)	(0.202)
GPA (9th Grade)	573683	2.896	(0.836)	(0.749)	(0.581)
In 10th grade on time	573963	0.899	(0.301)	(0.296)	(0.266)
GPA (10th Grade)	421872	2.76	(0.861)	(0.764)	(0.620)
Dropout (2005-2011 9th grade cohorts)	531920	0.043	(0.204)	(0.202)	(0.187)
Graduate (2005-2011 9th grade cohorts)	531920	0.824	(0.381)	(0.374)	(0.344)
Take SAT (2006-2011 9th grade cohorts)	472480	0.477	(0.499)	(0.479)	(0.410)
SAT Total Score	225684	1003.3	(189.0)	(165.9)	(123.7)
GPA at High-school Graduation	406826	2.809	(0.703)	(0.646)	(0.504)
Intend to attend 4yr college (2006-2011 9th grade cohorts)	472480	0.3506	(0.477)	(0.422)	(0.349)

Note: The sample uses data on all public school students in ninth grade in North Carolina between 2005 and 2012. The population is all students who took the English (English I) and math (algebra I, geometry, or algebra II) courses during ninth grade and can be linked to their classroom teachers. Incoming math scores and reading scores are standardized to be mean zero, unit variance for all takers in that year.

a. Test scores in the sample are higher than average because the ninth graders successfully matched to their classroom teacher are slightly higher achieving on average. Also, test scores in seventh and eighth grades are higher than the average because (a) the sample is based on those higher achievers who remained in school through ninth grade, and (b) I use the most recent eighth or seventh grade score prior to ninth grade, which will tend to be higher for repeaters.

b. GPA in eighth grade is only observed for high school courses taken while in eighth grade.

TABLE 2
PREDICTING LONG-RUN OUTCOMES USING NINTH-GRADE SKILL MEASURES

	1	2	3	4	5	6	7
	Dataset: NCERDC Micro Data						
	Main Longer-run Outcomes				Additional Outcomes		
	Drop out	Graduate	Drop out	Graduate	High-school GPA at Graduation	Take SAT	Intend to Attend 4-year College
Grade Point Average (9th grade)	-0.0353**	0.0933**					
	[0.000760]	[0.00126]					
Log of # Absences+1 (9th grade)	0.00635**	-0.0198**					
	[0.000317]	[0.000552]					
Suspended (9th grade)	0.0177**	-0.0503**					
	[0.00225]	[0.00339]					
On time in 10th grade	-0.0761**	0.337**					
	[0.00188]	[0.00301]					
Math z-score (9th grade)	-0.00427**	0.00691**					
	[0.000443]	[0.000794]					
English z-score (9th grade)	-0.00539**	0.00503**					
	[0.000659]	[0.00112]					
Average Test Scores: z-score [a]			-0.0133**	0.0186**	0.151**	0.0465**	0.0358**
			[0.000747]	[0.00113]	[0.00151]	[0.00128]	[0.00125]
Behavior index: z-score			-0.0524**	0.158**	0.345**	0.130**	0.09312**
			[0.000588]	[0.000781]	[0.00128]	[0.00073]	[0.00069]
Observations	439,284	439,284	527,571	527,571	403,672	468,015	468,015

Robust standard errors in brackets.
In addition to school fixed effects and year fixed effects, all models include controls for student gender, ethnicity, parental education, a cubic function of Math and Reading test scores in seventh and eighth grade, suspension in seventh and eighth grade, days absent in seventh and eighth grade, GPA in eighth grade [for high-school courses only], and whether the student had repeated seventh or eighth grade. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.
a. Where only one test-score is available, the average is the single available test score. As such, there are more observations with average scores that those with both English and Math scores.
** p<0.01, * p<0.05, + p<0.1

TABLE 3
COVARIANCE-BASED ESTIMATES OF THE VARIABILITY OF TEACHER VALUE-ADDED

	All Teachers							
	English Score	Math Score	Suspended	Absences [a]	9^{th} Grade GPA	In 10^{th} on time	Behaviors Index	10^{th} Grade GPA
English Teachers SD	0.0301	0.0292	0.0104	0.0434	0.0415	0.0212	0.0552	0.0360
Math Teachers SD	0.0204	0.0844	0.0121	0.0001	0.0632	0.0264	0.0801	0.0501
All Teachers SD	0.018	0.0751	0.0108	0.02839	0.0446	0.0247	0.0769	0.0315

Note: The estimated standard deviations are the square root of the estimated covariances in mean residuals from equation [9] across classrooms for the same teacher. Specifically, I pair each classroom with a randomly chosen different classroom for the same teacher and estimate the covariance. I replicate this 200 times and take the median estimated covariance as the parameter estimate. I then take the square root of this estimated covariance parameter as the estimated standard deviation of teacher value-added.

a. Absences refers to the natural log of the number of absences plus one.

TABLE 4
CORRELATIONS BETWEEN VALUE-ADDED ESTIMATES WITHIN THE SAME TEACHER

	Teacher Value-Added: Test Score	Teacher Value-Added: Suspended	Teacher Value-Added: Absences [a]	Teacher Value-Added: GPA	Teacher Value-Added: In 10th Grade On time	Teacher Value-Added: Behavior index	Teacher Effect: 10th Grade GPA
Teacher Value-Added: Test Score	1						
Teacher Value-Added: Suspended	-0.0726	1					
Teacher Value-Added: Absences [a]	-0.0390	0.0983	1				
Teacher Value-Added: GPA	0.2266	-0.1454	-0.0863	1			
Teacher Value-Added: In 10th Grade On time	0.1610	-0.1185	-0.0642	0.3822	1		
Teacher Value-Added: Behavior index	0.1494	-0.3325	-0.4461	0.5716	0.5454	1	
Teacher Value-Added: 10th Grade GPA	0.1147	-0.0449	-0.0601	0.3471	0.0973	0.2320	1

Note: This table reports the estimated two-way correlation coefficient between the estimated teacher value-added ($\hat{\mu}_{zjt}$) on each skill measure and each other skill measure.

a. Absences refers to the natural log of the number of absences plus one.

TABLE 5

EFFECTS OF TEACHER VALUE-ADDED ON SHORT-RUN SKILL MEASURES

	1	2	3	4	5	6	7	8	9	10	11
	Student Test-score in 9th Grade					Student Behaviors in 9th Grade			Student GPA in 10th Grade		
Teacher Value-Added: 9th Grade Test Score	0.0685**		0.0690**		0.0685**	0.0074**		0.0061**	0.0042**		0.0038**
	[0.0028]		[0.0028]		[0.00281]	[0.0016]		[0.0016]	[0.0012]		[0.0013]
Teacher Value-Added: 9th Grade Behaviors		0.0274*	-0.0214+				0.0579**	0.0536**			
		[0.0124]	[0.0128]				[0.0106]	[0.0107]			
Teacher Value-Added: 10th Grade GPA				0.0987**	0.0012					0.0357*	0.0298*
				[0.0230]	[0.0206]					[0.0147]	[0.0149]
School-Track Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	942,291	942,291	942,291	942,291	942,291	942,291	942,291	942,291	728,529	728,529	728,529

Note: Robust standard errors in brackets adjusted for two-way clustering at the teacher level and student level.

These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.

** p<0.01, * p<0.05, + p<0.1

TABLE 6

EFFECTS OF TEACHER VALUE-ADDED ON HIGH-SCHOOL COMPLETION

	1	2	3	4	5	6	7	8	9	10
	Student: Graduate High School					Student: Dropout of High School				
	Linear Probability Model			Conditional Logit [b]		Linear Probability Model			Conditional Logit [a,b]	
Teacher Value-Added: 9th Grade Test Score	0.0015**	0.0012*	0.0013*	0.0088	0.01	-0.0004	-0.0003	-0.0004	-0.0055	-0.0084
	[0.0005]	[0.00054]	[0.0005]	[0.0064]	[0.0064]	[0.0003]	[0.0003]	[0.0003]	[0.0099]	[0.0101]
				(0.002)	(0.0023)				(-0.0012)	(-0.0018)
Teacher Value-Added: 9th Grade Behaviors		0.0146**		0.1442**			-0.0041*		-0.128*	
		[0.00319]		[0.0343]			[0.0019]		[0.0583]	
				(0.0331)					(-0.0271)	
Teacher Value-Added: 10th Grade GPA			0.0146**		0.162**			-0.0031		-0.0618
			[0.0056]		[0.0637]			[0.0031]		[0.0996]
					(0.0375)					(-0.012)
% Increase in explained variance		305%	97%				326%	59%		
School-Track Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	891,868	891,868	891,868	579,512	579,512	891,868	891,868	891,868	570,390	570,390

Note: Robust standard errors in brackets adjusted for two-way clustering at both the teacher level and student level.

Standard errors are clustered at the teacher level for the conditional logit models. Implied average marginal effects from the conditional logit model are in parentheses.

These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.

To compute the increase in the explained variance, I compute the variance of the fitted values for each teacher in models without the behaviors value-added i.e. $a = var[\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt})]$, and in models with value-added on both i.e., $b = var[\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \hat{\delta}_2 \cdot (\varrho_2 \hat{\mu}_{behavior,jt})]$. The percentage increase in explained variability from also including value-added on behaviors (versus test-score value-added alone) is $100 \times ((\mathrm{b} \div \mathrm{a}) - 1)$.

a. Conditional logit models will not converge using the full sample. Tracks with a large number of observations led to a lack of convergence. As such, the conditional logit models are estimated in track cells with 500 or fewer observations. This accounts for roughly 80 percent of the data.

b. Note that conditional logit models drop observations in tracks with no variance. As such the number of observations used in the conditional logit model differs from that in the linear probability models.

** p<0.01, * p<0.05, + p<0.1

TABLE 7

EFFECTS OF TEACHER VALUE-ADDED ON VARIOUS LONG-TERM OUTCOMES

	1	2	3	4	5	9	7	8
	Enrolled in 10th Grade	10th Grade GPA	Dropout of School	Graduate High School	Take the SAT	Total SAT	Intend to Attend 4-year College [a]	GPA in 12th grade
Teacher Value-Added: 9th Grade Test Score	0.000836+	0.00389**	-0.000315	0.00118*	0.00114+	0.596*	0.00115	0.00109
	[0.000498]	[0.00120]	[0.000290]	[0.000546]	[0.000686]	[0.274]	[0.000801]	[0.000873]
Teacher Value-Added: 9th Grade Behaviors	0.0204**	0.0130+	-0.00407*	0.0146**	0.0116**	-0.232	0.01031*	0.0214**
	[0.00318]	[0.00786]	[0.00192]	[0.00319]	[0.00378]	[1.765]	[0.00442]	[0.00566]
% increase in explained variance	793%	33%	326%	305%	228%	0.10%	193%	607%
Teacher Value-Added: 9th Grade Test Score	0.00116*	0.00375**	-0.000363	0.00130*	0.00129+	0.596*	0.00128	0.00129
	[0.000521]	[0.00119]	[0.000292]	[0.000556]	[0.000696]	[0.275]	[0.000811]	[0.000877]
Teacher Value-Added: 10th Grade GPA	0.00974*	0.0298*	-0.00402	0.0146**	0.00796	-0.319	0.01142	0.0190*
	[0.00495]	[0.0149]	[0.00305]	[0.00565]	[0.00701]	[3.008]	[0.00865]	[0.00905]
% increase in explained variance	57%	54%	59%	97%	35%	0.25%	88%	151%
Observations	942,291	728,529	891,868	891,868	789,627	401,744	789,627	701,813

Note: Robust standard errors in brackets are adjusted for clustering at both the teacher and student level.

These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.

To compute the increase in the variance explained, I compute the variance of the fitted values for each teacher in models without the value-added on behaviors (i.e. $a = var(\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}))$, and in models with teacher effects on both (i.e. $b = var[\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \hat{\delta}_2 \cdot (\varrho_2 \hat{\mu}_{behavior,jt})]$. The percentage increase in explained variability from also including value-added on behaviors (versus test-score value-added alone) is $100 \times ((b \div a) - 1)$.

a. Note that intentions to attend college are only available for the 2006 through 2011 cohorts.

** p<0.01, * p<0.05, + p<0.1

REFERENCES

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

Alexander, K. L., D. R. Entwisle, and M. S. Thompson. 1987. "School Performance, Status Relations, and the Structure of Sentiment: Bringing the Teacher Back In." *American Sociological Review* 52: 665–82.

Allen, J. P., R. C. Pianta, A. Gregory, A. Y. Mikami, and J. Lun. 2011. "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement." *Science* 333: 1034–37.

Bacher-Hicks, A., Thomas J. Kane, and Douglas O. Staiger. 2015. "Validating Teacher Effect Estimates Using Changes in Teacher Assignment in Los Angeles." Working paper, Harvard University.

Barbaranelli, C., G. V. Caprara, A. Rabasca, and C. Pastorelli. 2003. "A Questionnaire for Measuring the Big Five in Late Childhood." *Personality and Individual Differences* 34(4): 645–64.

Bertrand, Marianne, and Jessica Pan. 2013. "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics* 5(1): 32–64.

Booker, K., T. R. Sass, B. Gill, and R. Zimmer. 2011. "The Effect of Charter High Schools on Educational Attainment." *Journal of Labor Economics* 29(2): 377–415.

Borghans, L., B. T. Weel, and B. A. Weinberg. 2008. "Interpersonal Styles and Labor Market Outcomes." *Journal of Human Resources* 43(4): 815–58.

Brookhart, S. M. 1993. "Teachers' Grading Practices: Meaning and Values." *Journal of Educational Measurement* 30(2): 123–42.

Cameron, Colin, Jonah Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multi-Way Clustering." *Journal of Business and Economic Statistics*, vol 21, No.2 pp238-249.

Carell, Scott E. and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3): 409–432.

Carneiro, P., C. Crawford, and A. Goodman. 2007. "The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes." CEE Discussion Paper 0092, Centre for the Economics of Education, London School of Economics and Political Science.

Chamberlain, Gary. 2013. "Predictive Effects of Teachers and Schools on Test Scores, College Attendance, and Earnings." Proceedings of the National Academy of Science (October 7). doi:10.1073/pnas.1315746110.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593–632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633–79.

Cunha, Flavio, and James J. Heckman. 2008. "Noncognitive Skills and Their Development." *Journal of Human Resources* 43(4): 738–82.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78(3): 883–931.

Deming, D. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3): 111–34.

Deming, D. 2011. "Better Schools, Less Crime?" *Quarterly Journal of Economics* 126(4): 2063–115.

Douglass, Harl R. 1958. "What Is a Good Teacher?" *High School Journal* 41(4): 110–13.

Downey, D., and P. Shana. 2004. "When Race Matters: Teachers' Evaluations of Students' Classroom Behavior." *Sociology of Education* 77: 267–82.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92(6): 1087–101.

Ehrenberg, R. G., D. D. Goldhaber, and D. J. Brewer. 1995. "Do Teachers' Race, Gender, and Ethnicity

Matter? Evidence from NELS88." *Industrial and Labor Relations Review* 48: 547–61.
Figlio, David N., Morton O. Schapiro, and Kevin B. Soter. 2015. "Are Tenure Track Professors Better Teachers?" *Review of Economics and Statistics* 97(4): 715–24.
Fredriksson, P., B. Ockert, and H. Oosterbeek. 2013. "Long-Term Effects of Class Size." *Quarterly Journal of Economics* 128 (1): 249-285.
Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. “Identifying Effective Teachers Using Performance on the Job,” Hamilton Project White Paper 2006-01.
Greene, William. 2002. *Econometric Analysis*. 5th ed. Upper Saddle River, New Jersey: Prentice Hall.
Harris, D., and A. Anderson. 2012. "Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence From Middle School Teachers." Panel Paper, Association of Public Policy Analysis and Management.
Heckman, James, (1981a), Heterogeneity and State Dependence, p. 91-140 in , Studies in Labor Markets, National Bureau of Economic Research, Inc.
Heckman, James (1981b), “Statistical Models for Discrete Panel Data” published in Structural Analysis of Discrete Data and Econometric Applications, Edited by Charles F. Manski and Daniel L. McFadden, Cambridge: The MIT Press, 1981
Heckman, James J. 1999. "Policies to Foster Human Capital." NBER Working Paper 7288, National Bureau of Economic Research.
Heckman, James J., John Eric Humphries, and Gregory Veramendi. 2016. "Dynamic Treatment Effects." *Journal of Econometrics* 191(2): 276–92.
Heckman, James J., and T. Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19(4), 451–64.
Heckman, James J., Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103(6): 2052–86.
Heckman, James J., and Y. Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review* 91(2): 145–49.
Heckman, James J., J. Stixrud, and S. Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24(3): 411–82.
Howley, A., P. S. Kusimo, and L. Parrott. 2000. "Grading and the Ethos of Effort." *Learning Environments Research* 3: 229–46.
Jackson, C. Kirabo. 2013. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers". *Review of Economics and Statistics* 95: 1096–116.
Jackson, C. Kirabo. 2014. "Teacher Quality at the High-School Level: The Importance of Accounting for Tracks." *Journal of Labor Economics* 32(4), 32 (4), 645–684.
Jackson, C. Kirabo., and E. Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics* 1(4): 85–108.
Jackson, Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher Related Policies." *Annual Review of Economics,* Vol. 6: 801-825.
Jacob, Brian, Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources* 45(4): 915–43.
Jennings, J. L., and T. A. DiPrete. 2010. "Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology of Education* 83(2): 135–59.
John, O., A. Caspi, R. Robins, T. Moffit, and M. Stouthamer-Loeber. 1994. "The “Little Five”: Exploring the Nomological Network of the Five-Index Model of Personality in Adolescent Boys". *Child Development* 65: 160–78.
Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. “Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.” MET Project Research Paper, Bill & Melinda Gates Foundation.
Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement:

An Experimental Evaluation." NBER Working Paper 14607, National Bureau of Economic Research.

Kautz, Tim, and Wladimir Zanoni. 2014. "Measuring and Fostering Non-Cognitive Skills in Adolescence: Evidence from Chicago Public Schools and the OneGoal Program." University of Chicago. http://home.uchicago.edu/~tkautz/OneGoal_TEXT.pdf.

Koedel, C. 2008. "Teacher Quality and Dropout Outcomes in a Large, Urban School District." *Journal of Urban Economics* 64(3): 560–72.

Lee, C. D. 2007. *The Role of Culture in Academic Literacies: Conducting Our Blooming in the Midst of the Whirlwind*. New York: Teachers College Press.

Lindqvist, E., and R. Vestman. 2011. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics* 3(1): 101–28.

Lleras, Christy. 2008. "Do Skills and Behaviors in High School Matter? The Contribution of Noncognitive Factors in Explaining Differences in Educational Attainment and Earnings." *Social Science Research* 37: 888–902.

Lounsbury, J. W., R. P. Steel, J. M. Loveland, and L. W. Gibson. 2004. "An Investigation of Personality Traits in Relation to Adolescent School Absenteeism." *Journal of Youth and Adolescence* 33(5): 457–66.

Lucas, S. R., and M. Berends. 2002. "Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking." *Sociology of Education* 75(4): 328–348.

Mansfield, R. 2012. "Teacher Quality and Student Inequality." Working paper, Cornell University.

Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood. 2013. "A Composite Estimator of Effective Teaching." MET Project Research Paper, Bill & Melinda Gates Foundation.

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.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417–58.

Rothstein, J. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, 125 (1): 175-214.

Sadker, D. M., and K. Zittleman. 2006. *Teachers, Schools and Society: A Brief Introduction to Education.* New York: McGraw-Hill.

Siskin, Leslie. 1991. "Departments As Different Worlds: Subject Subcultures In Secondary Schools" *Educational Administration Quarterly*, May: 124–60. http://www.academia.edu/335979/Departments_As_Different_Worlds_Subject_Subcultures_In_Secondary_Schools.

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *American Economic Review* 102(7): 3628–51.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113: F3–F33.

Tyler, John H. and Magnus Lofstrom. 2009. “Finishing High School: Alternative Pathways and Dropout Recovery.” *Future of Children* 19(1): 77–103.

United States Department of Labor, Bureau of Labor Statistics. 2016. "Employment Projections: Earnings and Unemployment Rates by Educational Attainment, 2015." Accessed February 12. http://www.bls.gov/emp/ep_chart_001.htm.

Waddell, G. 2006. "Labor-Market Consequences of Poor Attitude and Low Self-Esteem in Youth." *Economic Inquiry* 44(1): 69–97.

Appendix A

Matching Teachers to Students

The North Carolina Education Research Data Center (NCERDC) data contains End of Course (EOC) files with student 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. 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 is). 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). Following Mansfield (2012), I group students into classrooms based on the unique combination of class period, course type, subject code, test date, school code, and teacher ID code. I then compute classroom-level totals for 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 in the course), 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, geometry, algebra II or English I teacher are automatically linked to the teacher ID from the SAR files. These are perfectly matched. Matched classes are set aside.
(2) Classes that match exactly on all classroom characteristics and the teacher ID are deemed matches. These are deemed perfectly matched. Matched classes are set aside.
(3) Compute a score for each potential match (the sum of the squared difference between each observed classroom characteristic for classrooms in the same school in the same year in the same subject, and infinity otherwise) in the SAR file and the EOC data. Find the best match in the SAR file for each EOC classroom. If the best match also matches in the teacher ID, a match is made. These are deemed imperfectly matched. Matched classes are set aside.
(4) Find the best match (based on the score) in the SAR file for each EOC classroom. If the SAR classroom is also the best match in the EOC classroom for the SAR class, a match is made. These are deemed imperfectly matched. Matched classes are set aside.
(5) Repeat step 3 and 4 until no more-high quality matches can be made.

This procedure leads to a matching of 90 percent of English classrooms and 83 percent of math classrooms with ninth graders in the testing file. Results are similar when using cases in which the matching is exact, so error due to the fuzzy matching algorithm does not generate any of the empirical findings.

Appendix B

Correlations Between Individual Behaviors and Skill Measures

The correlations among the ninth-grade outcomes (or skill measures) are presented in table B1. They reveal some interesting patterns. The first pattern is that test scores are relatively strongly correlated both with each other and with grade point average (correlation≈0.55) but are weakly correlated with other non-test-score outcomes. Specifically, the correlations between the natural log of absences (note: 1 is added to absences before taking logs so that zeros are not dropped) is about -0.16 for both math and English test scores, and the correlations between being suspended are about -0.13 for both math and English test scores. While slightly higher, the correlation between on-time progression to tenth grade (i.e. being a tenth grader the following year) and test scores is only 0.28. This reveals that while students who tend to have better test-score performance also tend to have better non-test-score outcomes, the ability to predict non-test-score outcomes based on test scores is relatively limited. Simply put, students who score well on standardized tests are not necessarily those who are well-adjusted, and many students who are not well-behaved score well on standardized tests. Indeed, the implied R^2s from the correlations in table B1 indicate that test scores predict less than five percent of the variation in absences and suspensions, less than 10 percent of the variation in on-time grade progression, and about one-third of the variation in GPA. Because these outcomes are interesting in their own right, test scores may not measure *overall* educational well-being.

The second notable pattern is that many behaviors are more highly correlated with each other than with scores. For example, the correlations between suspensions and test scores are smaller than those between suspensions and all other outcomes. Similarly, the correlations between absences and test scores are smaller than those between absences and other outcomes. The third notable pattern is that GPA is relatively well correlated with both test-score and non-test-score outcomes. This is consistent with research (Howley, Kusimo, and Parrott 2000; Brookhart 1993) finding that most teachers base their grading on some combination of student product (exam scores, final reports, etc.), student process (effort, class behavior, punctuality, etc.) and student progress—so that grades reflect a combination of cognitive and noncognitive skills.

The patterns suggest that the outcomes can be put into three categories: academic aptitude variables (math and English test scores), behavioral variables (absences and suspensions) and those that reflect a combination of aptitude and behaviors (on-time grade progression and GPA). It seems likely that each of these three groups of variables may reflect a somewhat different combination of cognitive and noncognitive skills. If teachers improve student outcomes by improving both cognitive and noncognitive skills, their value-added on a combination of these outcomes should better predict their impact on longer-run outcomes than using their test score value-added alone.

TABLE B1

RAW TWO-WAY CORRELATION COEFFICIENTS BETWEEN OUTCOMES

	Log of # Days Absent +1	Suspended	Grade Point Average	In 10^{th} grade on time	Math Score 9^{th} Grade	English Score 9^{th} Grade	Behavior Index	Test-Score Index
Log of # Days Absent +1	1							
Suspended	0.183	1						
Grade Point Average	-0.298	-0.205	1					
In 10^{th} grade on time	-0.187	-0.151	0.428	1				
Math Score 9^{th} Grade	-0.163	-0.122	0.552	0.266	1			
English Score 9^{th} Grade	-0.151	-0.148	0.594	0.290	0.538	1		
Behavior Index							1	
Test-score Index							0.5593	1

The dataset used to compute these correlations includes one observation per student. There are 573,683 student observations. The behavior index was uncovered using principal component analysis and is a linear combination of all the short-run non-test-score outcomes. Specifically, this noncognitive index is 0.38(GPA) + 0.31(in tenth grade) − 0.15(suspended) − 0.21(log of 1 + absences). The weighted average is then standardized to have a mean of zero and unit variance. The test-score index is the equal weight average of the test-score outcomes. It is also standardized to unit variance with a mean of zero.

Appendix C

Analysis of the NELS-88 Data

To ensure that the patterns in table 2 are not specific to North Carolina, I also employ data from the National Educational Longitudinal Survey of 1988 (NELS-88). The NELS-88 is a nationally representative sample of respondents who were eighth-graders in 1988. Table C1 presents the same models using the NELS-88 data. I predict longer-term outcomes as a function of the same behavioral outcomes and test-score variables used in the NCERDC data. For both dropout and high-school graduation, increases in the behavior index are associated with large improvements in longer-run outcomes conditional on test scores. Looking at college going, a 1σ increase in the test-score index (the average of math and English scores as in table 2) is associated with a 5.2 percentage point increase in college-going while a 1σ increase in the behavior index is associated with a 9.6 percentage point increase.

The NELS-88 data also include long-term outcomes, collected when the respondents were 25 years old. These allow one to see how this behavior index (based on eighth-grade outcomes) predicts being arrested (or having a close friend who was arrested), employment, and labor market earnings, conditional on eighth-grade test scores. The results show that test scores are actually positively associated with being arrested (conditional on all the covariates), but a 1σ increase in the behavior index is associated with a 5.6 percentage point decrease in being arrested (or having a close friend who was arrested). Looking to labor market outcomes, both test scores and the behavior index predict employment in the labor market and earnings. Specifically, a 1σ increase in test scores is associated with a 1.3 percentage point increase in working, while a 1σ increase in the behavior index is associated with a 2 percentage point increase. Finally, conditional on having any earnings, a 1σ increase in test scores is associated with 14.4 percent higher earnings while a 1σ increase in the behavior index is associated with 24.6 percent higher earnings.

In recent findings, both Lindqvist and Vestman (2011) and Heckman, Stixrud, and Urzua (2006) find that noncognitive ability is particularly important at the lower end of the earnings distribution. In particular, using high-quality detailed psychometric measures of noncognitive skills, Lindqvist and Vestman (2011) find that in bottom 25th percent of the earings distribution, a 1σ increas in noncognitve skills is associated with about 25 percent higher earnings, while for the top 25th percent, this is about 8 percent. Looking at test scores, they find that a 1σ increase in test scores is associated with about 9 percent higher earnings throughout the earnings distribution. Insofar as the behavior index captures noncognitive skills, one would expect this to be the case for this index also. To test this, I estimate unconditional quantile regressions to obtain the marginal effect on log wages at different points in the earnings distribution. The results (table C2) show that at the 90th percentile through the 75th percentile of the earnings distribution, a 1σ increase in test scores and the behavior index is associated with a very similar increase of between 5 and 6 percent higher earnings. However, at the median, the behavior index is more important; the marginal effect of a 1σ increase in test scores and the behavior index are 3.2 percent and 10 percent higher earnings, respectively. At the 25th percentile, this difference is even more pronounced. A 1σ increase in test scores is associated with 3.1 percent higher earnings while a 1σ increase in the behavior index is associated with 23 percent higher earnings. These findings are remarkably similar to those presented by Lindqvist and Vestman (2011) that use psychometric measures of noncognitive skills, suggesting that this index is a reasonable proxy for noncognitive ability.

TABLE C1

RELATIONSHIP BETWEEN SHORT-RUN SKILL MEASURES AND LONGER-RUN OUTCOMES

	1	2	3	4	5	6
	Dataset: National Educational Longitudinal Survey 1988					
	Dropout	Graduate	College (by age 25)	Arrests (by age 25)	Working (at age 25)	Log Income (at age 25)
Test-score index: z-score	-0.00923**	0.00304	0.0522**	0.0151*	0.0131**	0.144**
	[0.00256]	[0.00407]	[0.00575]	[0.00610]	[0.00506]	[0.0506]
Behavior index: z-score	-0.0482**	0.0933**	0.0955**	-0.0559**	0.0200**	0.246**
	[0.00339]	[0.00442]	[0.00533]	[0.00566]	[0.00470]	[0.0467]
School Fixed Effects	Y	Y	Y	Y	Y	Y
Covariates	Y	Y	Y	Y	Y	Y
Observations	10,792	10,792	10,792	10,792	10,792	10,792

Robust standard errors in brackets.

All models control for ethnicity, gender, family income, family size, and school fixed effects.
** p<0.01, * p<0.05, + p<0.1

TABLE C2

EFFECT OF TEST SCORES AND THE BEHAVIOR INDEX IN EIGHTH GRADE ON ADULT EARNINGS AT DIFFERENT PERCENTILES (NELS-88)

	Natural log of Income (age 25): Conditional on Working			
Percentile	25th	50th	75th	90th
Test-score index: z-score	0.00312	0.0318**	0.0495**	0.0582**
	[0.0511]	[0.00939]	[0.00691]	[0.00866]
Behavior index: z-score	0.233**	0.100**	0.0679**	0.0509**
	[0.0467]	[0.00858]	[0.00632]	[0.00791]
School Fixed Effects	Y	Y	Y	Y
Covariates	Y	Y	Y	Y
Observations	10,792	10,792	10,792	10,792

Standard errors in brackets.
All models control for ethnicity, gender, family income, family size, and school fixed effects.
** p<0.01, * p<0.05, + p<0.1

Appendix D

Formal Proofs for Section III

The Production Function Yielding the Additively Separable Model

For ease of exposition, the model presented in section III is one in which student and teacher contributions to outcomes are additive. This is typical of the teacher value-added literature. However, appendix D outlines the underlying Cobb-Douglas production function that gives rise to the simple additive model outlined in the main text. All the parameters presented in the simplified model in Section III are derived here. Note that many of the parameters are first presented in levels here, and then later presented in logs. In the main text these parameters are introduced in log form.

Production of Student Skills: Prior to ninth grade, each student *i* has a stock of cognitive and noncognitive abilities described by vector $\mathrm{N}_i = (\mathrm{N}_{ci}, \mathrm{N}_{ni})^T$, where the subscripts *c* and *n* denote the cognitive and noncognitive dimensions, respectively.[1] This stock reflects an initial endowment and the cumulative effect of all school and parental inputs on student incoming abilities.

During ninth grade, students take classes in many subjects (e.g. math, English, sciences, social studies, etc.), so that students are exposed to up to seven different teachers at each school across the subjects $d \in \{1,2,3, \dots, 7\}$. Each ninth-grade teacher *j* in subject *d* has a positive quality vector $\Omega_{jd} = (\Omega_{cjd}, \Omega_{njd})^T$ describing teacher *j's* capacity to increase each of the two dimensions of student ability during ninth grade. **Note** that teacher *j* in subject *d* is a different teacher from teacher *j* in subject *d'*, such that the unique combination of *d* and *j* defines an individual teacher. Student ability at the end of ninth grade is a function of student ability at the *beginning* of the year, and the contribution of each teacher *j* across all subjects *d*. Production of student ability in each dimension at the end of ninth grade is Cobb-Douglas as below in [D1].

$$\text{[D1]} \qquad e^{\alpha_{ij}} = \begin{pmatrix} e^{\alpha_{cij}} \\ e^{\alpha_{nij}} \end{pmatrix} = \begin{pmatrix} (\mathrm{N}_{ci})^{\mathrm{p}_0} (\Omega_{cj1})^{\mathrm{p}_{ci1}} (\Omega_{cj2})^{\mathrm{p}_{ci2}} \dots (\Omega_{cj7})^{\mathrm{p}_{ci7}} \\ (\mathrm{N}_{ni})^{\mathrm{p}_0} (\Omega_{nj1})^{\mathrm{p}_{ni1}} (\Omega_{nj2})^{\mathrm{p}_{ni2}} \dots (\Omega_{nj7})^{\mathrm{p}_{ni7}} \end{pmatrix}.$$

The production function parameters p_{cid} and p_{nid} connote the relative importance of each teacher input in producing each of the two dimensions of ability. The *i* subscripts on the production function parameters connote that the end of year ability of each student *i* may be differentially responsive to the quality of teacher *j* in subject *d*. The natural log of student ability at the end of the year is then the contribution of incoming ability plus the sum of the contributions of each of the student's teachers in each subject as in [D2].

$$\text{[D2]} \quad \alpha_{ij} = \mathrm{p}_0 \log(\mathrm{N}_i) + \textstyle\sum_{d=1}^{7} \mathrm{p}_{id} \cdot \log(\Omega_{jd}) \equiv \begin{pmatrix} \mathrm{p}_0 \log(\mathrm{N}_{ci}) + \sum_{d=1}^{7} \mathrm{p}_{cid} \cdot \log(\Omega_{cjd}) \\ \mathrm{p}_0 \log(\mathrm{N}_{ni}) + \sum_{d=1}^{7} \mathrm{p}_{nid} \cdot \log(\Omega_{njd}) \end{pmatrix}.$$

To simplify notation, let $\upsilon_i = \mathrm{p}_0 \log(\mathrm{N}_i)$ denote the contribution of incoming student ability to ability at the end of ninth grade, and let $\omega_{jd} = \log(\Omega_{jd})$ denote the teacher ability vector. Because students are differentially responsive to teacher ability, let $D_i = \begin{bmatrix} \mathrm{p}_{cid} & 0 \\ 0 & \mathrm{p}_{nid} \end{bmatrix} \equiv \begin{bmatrix} D_{ci} & 0 \\ 0 & D_{ni} \end{bmatrix}$, define the student responsiveness to teacher ability such that $\omega_{ijd} = D_i \omega_{jd}$ is the contribution of teacher *j* in subject *d* to the two-dimensional ability of student *i* at the end of ninth grade. That is,

[1] Students may possess many types of cognitive and noncognitive skills. The key point is that the extension relaxes the assumption that students are either high- or low-skilled, and permits the more realistic scenario in which students may be highly skilled in certain dimensions but deficient in other dimensions of skill.

ω_{ijd} is the "effective" quality of teacher j in subject d for student i and is the student "responsiveness" matrix (D_i) times the underlying quality vector of teacher j (ω_{jd}) .[2] To simplify notation further, let φ_{i-j}, denote the sum of the contributions of the other teachers (i.e. teachers in all *other* subjects $d' \neq d$) to the two-dimensional ability of student *i* at the end of ninth grade, such that $\varphi_{i-j} = \begin{pmatrix} \sum_{d' \neq d} p_{cid'} \cdot \log(\Omega_{cjd'}) \\ \sum_{d' \neq d} p_{nid'} \cdot \log(\Omega_{njd'}) \end{pmatrix}$. Ability at the end of ninth grade of student *i* with teacher *j* in subject *d* can be represented by the vector in [D3].

[D3] $$\alpha_{ijd} = \upsilon_i + D_i\omega_{jd} + \varphi_{i-j} = \upsilon_i + \omega_{ijd} + \varphi_{i-j}$$

All analyses are performed within-subject (i.e. I only compare the outcomes of students across teacher in the same subject). As such, for parsimony, I drop the subscript *d* yielding the additively separable model in [D4] below.

[D4] $$\alpha_{ij} = \upsilon_i + \omega_{ij} + \varphi_{i-j}$$

<u>*Formal Proof of Claim in Section III*</u>

<u>Claim:</u> *Teacher value-added on y_2 (θ_{2j}) will increase the explained teacher-level variation in the longer-run outcome iff $cov(\ddot{\theta}_{lj}, \ddot{\theta}_{2j}) \neq 0$.*

Proof: The variance of the average longer-run effect explained by value-added on skill measure 1 (i.e. test scores) in a linear regression model is simply $A \equiv var(\gamma_{1a}\theta_{1j})$, where γ_{1a} is the coefficient of θ_{1j} in a simple linear regression predicting θ_{lj}. In a model with both value-added on skill measure 1 (test scores) and value-added on skill measure 2 (i.e. behaviors), the explained variance is $B \equiv var(\gamma_{1b}\theta_{1j} + \gamma_{2b}\theta_{2j})$, where γ_{1b} and γ_{2b} are the coefficients of θ_{1j} and θ_{2j} in a multivariate linear regression predicting θ_{lj}, respectively.

From Greene (2002, 30), $B \equiv var\left(\gamma_{1b}\theta_{1j} + \gamma_{2b}\theta_{2j}\right) = var(\gamma_{1a}\theta_{1j} + \gamma_{2a}\ddot{\theta}_{2j})$ where $\ddot{\theta}_{2j}$ is the residual of θ_{2i} (after removing the linear association with θ_{1j}), and γ_{2a} is the coefficient on $\ddot{\theta}_{2j}$ in predicting $\ddot{\theta}_{lj}$. Recall, $\ddot{\theta}_{lj}$ is the residual average effect on longer-run outcomes after removing the linear association with θ_{1j}. Because $\ddot{\theta}_{2j}$ is uncorrelated with θ_{1j} by construction, it follows that $B = A + (\gamma_{2a})^2 \times var(\ddot{\theta}_{2j})$. Given that $var\left(\ddot{\theta}_{2j}\right) > 0$, the explained variance will be greater with value-added on both skill measures than with only test-score value-added (i.e. B>A) if $\gamma_{2a} \neq 0$. Because $\gamma_{2a} = cov(f(\omega_j), g(\omega_j))/var(\omega_j)$, it follows that $\gamma_{2a} \neq 0$ *iff* $cov\left(f\left(\omega_j\right), g\left(\omega_j\right)\right) \equiv cov(\ddot{\theta}_{lj}, \ddot{\theta}_{2j}) \neq 0$.

[2] Note that $\omega_{ijd} = \left(p_{cid}.\omega_{cjd}, p_{nid}.\omega_{njd}\right)^T$ which is a two-dimensional student-specific teacher quality vector. This relaxes the commonly-made assumption that teacher effects are the same for all students (see Jackson, Rockoff, and Staiger 2014).

Appendix E

The Creation of Tracks

Even though schools may not have explicit labels for tracks, most practice de-facto tracking by placing students of differing levels of perceived ability into distinct groups of courses (Sadker and Zittleman, 2006; Lucas and Berends, 2002). While there are many courses that ninth-grade students can take (including special topics and reading groups), there are 10 academic courses that constitute two-thirds of all courses taken. They are listed in table E1. As highlighted in Jackson (2014) and Harris and Anderson (2012), it is not only the course that matters but also the level at which the student takes the course. As such, following Jackson (2014), a school track is the unique combination of the ten most common academic courses, the level of algebra I taken, and the level of English I taken, in a particular school. Defining tracks flexibly at the school by course-group by course level allows for different schools that have different selection models and treatments for each track. As such, only students at the same school who take the same academic courses, level of English I, and level of Algebra I are in the same school track. There are 31,610 tracks across the 955,678 student observations. Because many students pursue the same course of study, less than one percent of all students are in singleton tracks, 83 percent of students are in tracks with more than 20 students, and the median student is in a school track with 199 other students. Including indicators for each school track in a value-added model compares outcomes across teachers within groups of students *in the same track at the same school*. This removes the influence of both track-level treatments and selection to tracks on estimated teacher effects.

All inference is made within school tracks so that identification of teacher effects comes from two sources of variation: (1) comparisons of teachers at the same school teaching students in the same track at different points in time and (2) comparisons of teachers at the same school teaching students in the same track at the same time. To compare variation within school tracks during the same year to variation within school tracks across years (cohorts), I compute the number of teachers in each non-singleton school-track-year-cell for both math and English (table E2). About 59 and 63 percent of all school-track-year cells include one teacher in the English course and the math course, respectively. As such, much variation is based on comparing single teachers across cohorts within the same school track. Appendix G shows that results using variation within school-track-cohort cells are similar to those obtained using only variation entirely across cohorts within a school.

TABLE E1

MOST COMMON ACADEMIC COURSES

Academic course rank	Course Name	% of 9th graders taking
1	English I	94
2	World History	85
3	Earth Science	57
4	Algebra I	61
5	Geometry	22
6	Art I	16
7	Biology I	15
9	Algebra II	14
9	Basic Earth Science	13
10	Spanish I	13

Note: Algebra I includes Introduction to algebra

TABLE E2

DISTRIBUTION OF NUMBER OF TEACHERS IN EACH SCHOOL-TRACK-YEAR CELL

	Percent	
Number of Teachers in School-Track-Year Cell	English	Math
1	59.59	63.43
2	20.53	19.18
3	9.79	8.39
4	4.85	4.17
5	2.53	2.16
6	1.32	1.18
7+	1.4	1.5

Appendix F

Showing Effects of Teachers on Individual Behavioral Outcomes

To show that the relationships between longer-run outcomes and teacher value-added on the behavior index are not driven by any single behavior, I estimate a model similar to equation [12] but in which I use the estimated teacher value-added on the individual behaviors instead of using value-added on the aggregate behavior index. I also present results using value-added on a behavior index that is based only on absences, suspensions, and on-time grade progression (i.e., excluding GPA).

For both graduation and dropout outcomes (tables F1 and F2), teacher value-added on the index excluding GPA predict the long-term outcomes—showing that the GPA variable does not drive the results. One can also see that teacher value-added on absences, GPA, and on-time grade progression each independently predict teacher impacts on high-school graduation —showing that no single variable drives the results. Finally, teacher value-added on the behavior index that combines all the behaviors is more strongly (in a statistical sense) associated with improved longer-term outcomes than the value-added on each of the individual outcomes. This indicates that it is improvement in those skills common to *all the behaviors* that drives the results.

TABLE F1

EFFECT OF VALUE-ADDED ON HIGH-SCHOOL GRADUATION: FOR INDIVIDUAL BEHAVIORS

	1	2	3	4	5	6	7
	Graduate from High School						
Value-Added: Test Score	0.00118*	0.00136*	0.00149**	0.00147**	0.00128*	0.00140*	0.00130*
	[0.000546]	[0.000547]	[0.000548]	[0.000548]	[0.000560]	[0.000546]	[0.000556]
Value-Added: Behavior index	0.0146**						
	[0.00319]						
Value-Added: Behavior index w/o GPA		0.0553*					
		[0.0228]					
Value-Added: Suspended			-0.0444				
			[0.0352]				
Value-Added: Absences [a]				-0.0307+			
				[0.0172]			
Value-Added: GPA					0.00384*		
					[0.00182]		
Value-Added: In 10^{th} Grade on time						0.00912+	
						[0.00502]	
Value-Added: GPA in 10th Grade							0.0146**
							[0.00565]
Observations	891,868	891,868	891,868	891,868	891,868	891,868	891,868

Robust standard errors in brackets are adjusted for clustering at both the teacher and student level.
These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.

a. Absences refers to the natural log of the number of absences plus one.

** $p<0.01$, * $p<0.05$, + $p<0.1$

TABLE F2

EFFECT OF VALUE-ADDED ON DROPPING OUT OF HIGH SCHOOL: FOR INDIVIDUAL BEHAVIORS

	1	2	3	4	5	6	7
	Dropping Out of School						
Value-Added: Test Score	-0.000315	-0.00038	-0.000419	-0.000418	-0.000338	-0.000343	-0.000363
	[0.000290]	[0.000292]	[0.000290]	[0.000291]	[0.000294]	[0.000289]	[0.000292]
Value-Added: Behavior index	-0.00407*						
	[0.00192]						
Value-Added: Behavior index w/o GPA		-0.024+					
		[0.0127]					
Value-Added: Suspended			-0.0182				
			[0.0193]				
Value-Added: Absences [a]				-0.00577			
				[0.00954]			
Value-Added: GPA					-0.00114		
					[0.000978]		
Value-Added: In 10^{th} Grade on						-0.00462+	
						[0.00273]	
Value-Added: GPA in 10th Grade							-0.00307
							[0.00305]
Observations	891,868	891,868	891,868	891,868	891,868	891,868	891,868

Robust standard errors in brackets are adjusted for clustering at both the teacher and student level.

These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.

a. Absences refers to the natural log of the number of absences plus one.

** p<0.01, * p<0.05, + p<0.1

Appendix G

Testing the Identifying Assumptions

The main results in this paper rely on the validity of two identifying assumptions. Even though there is no dispositive way to prove the validity of these assumptions, the specification and falsification checks presented in this section provide empirical evidence that both identifying assumptions are likely satisfied.

Testing for bias due to Selection of Students to Teachers

The first identifying assumption is that conditional on controls for tracking and sorting there is no selection of students to ninth-grade teachers. If there is a subset of variables from X_{icjst}, say Z_{2icjst} conditional on which there is no sorting of students to teachers, one can test for selection on observables. Consider the following logic. Let Z_{1it} be all those variables included in X_{icjst}, but not in Z_{2icjst}. If conditional on Z_{2icjst} there is no relationship between estimated teacher value-added ($\hat{\mu}_{zjt}$) and Z_{1it}, it would suggest no selection of students to teachers on observables.

To present such evidence that the results are not driven by selection, I predict each outcome based on a linear regression of each outcome on seventh-grade math and reading scores, seventh-grade repetition, suspensions in seventh grade and absences in seventh grade, parental education, gender, and ethnicity. Specifically, where y_{zicjst} is high-school graduation and high-school dropout, Z_{1it} are the demographic variables and the seventh grade skill measures described above, I model y_{zicjst} as a function of Z_{1it} by logistic regression. I then take the predicted outcome from this regression $\tilde{y}_{zicjst}$.

I then regress predicted outcomes on the set of covariates Z_{2icjst} (which is all the covariates in X_{icjst} excluding those in Z_{1it}) and the leave-year-out teacher value-added as in [G1], where all variables are defined as previously and $\dot{\varepsilon}_{zicjst}$ is a random student-level error.

[G1] $$\tilde{y}_{zicjst} = \Pi_{2z} Z_{2icjst} + \xi_{z1} \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \xi_{z2} \cdot (\varrho_2 \hat{\mu}_{behavior,jt}) + \iota_d \sum_{d=1}^{4} I_d + \dot{\varepsilon}_{zicjst}$$

If the estimated value-added estimates were driven by positive selection to teachers, one might observe a positive relationship between value-added and the predicted outcomes. Results are in table G1. Columns 1 through 4 present the main results only conditional on Z_{2icjst} (i.e. excluding controls for seventh-grade skill measures and demographics). The results are almost identical to those in Table 6, which include the full set of controls, so that excluding the demographic variables and the seventh-grade skill measures has a negligible effect on the main results. This suggests little bias due to selection on observables. Testing this more directly, columns 5 through 8 show that value-added conditional on controls for tracking and eighth-grade skill measures are largely unrelated to predicted outcomes (i.e. largely unrelated to outcomes as predicted by parental education, gender, ethnicity, and seventh-grade skill measures -- all of which are strong predictors of the longer-run outcomes). For behaviors value-added, the points estimates are both small and indistinguishable from zero (at traditional levels of significance), indicating no selection of students to teachers behaviors value-added on *observables*. Though the *p*-value for test score value-added for predicted graduation is slightly smaller than 0.1, it is not significant at the 5 percent level, the point estimates are small, and the *p*-values for predicted dropout are large. Moreover, the point estimates are negative for *both* predicted dropout and predicted graduation, which is inconsistent with any systematic sorting of teachers. Taken together the results suggest that there is little to no selection of students to teachers on *observables*.

Even though the results thus far are reassuring regarding selection on observables, one may worry about selection on *unobservables*. To address this concern, I present results that rely on two distinct sources of variation and that rely on different identifying assumptions. Specifically, I first present a strategy that relies exclusively on *within* school-year variation in estimated teacher quality. Because selection to teachers occurs within ninth-grade cohorts within schools, models based on this variation are most susceptible to being biased by selection on *unobervables*. I then present an instrumental variables (IV) strategy that relies only on variation in average estimated teacher quality *across* cohorts entire ninth-grade cohorts within schools. The within-school-cohort strategy presented is robust to any school-level changes or polices that can impact outcomes, but is susceptible to bias due to selection within a school-cohort. Conversely, the across-cohort within-school instrumental variables strategy presented is robust to selection of students to teachers within a school-cohort, but is susceptible to bias due to school-level changes or polices that can impact outcomes. Because it is unlikely that both biases yield the same empirical patterns, if the results are similar across these two different strategies, it is implied that the estimated effects are real.

The within-school cohort identification strategy

From [10] and [11], the out of sample teacher value-added estimate is $\hat{\mu}_{zjt} = \bar{e}_{zj,-t}\lambda_{zj}$. Because $e_{zicjst} = \theta_{zj} + \varepsilon_{zcjst} + \varepsilon_{zicjst}$, the leave-year-out average of the student-level residuals for a given teacher can be written as $\bar{e}_{zj,-t} = \theta_{zj} + \bar{\varepsilon}_{zj,c\neq t} + \bar{\varepsilon}_{zj,i\neq t}$, where $\bar{\varepsilon}_{zj,c\neq t}$ is the average of the classroom-level shocks for teacher *j* excluding her classes in year *t*, and $\bar{\varepsilon}_{zj,i\neq t}$ is the average of the student-level unobserved characteristics for teacher *j* excluding her students in year *t*. This out-of-sample effect can be broken into three pieces; a piece that is due to real differences in teacher quality ($\lambda_{zj}\theta_{zj}$), a piece that is due to unobserved variability at the classroom level ($\lambda_{zj}\bar{\varepsilon}_{zj,c\neq t}$), and a piece that is due to the unobserved characteristics of the students assigned to teacher *j* within a cohort ($\lambda_{zj}\bar{\varepsilon}_{jz,i\neq t}$). As such, teachers have large positive estimated value-added either because they are truly good teachers (i.e. $\theta_{zj} > 0$) , because they happened to be lucky and have many positive classroom shocks in other years ($\bar{\varepsilon}_{zj,c\neq t} > 0$), or because they systematically tend to be assigned to students within a school and cohort with better unobserved characteristics ($\bar{\varepsilon}_{jz,i\neq t} > 0$).

To rely only on variation that occurs within cohorts within schools, I augment equation [12] to also include school-year fixed effects (τ_{st}) as in [G2] below.

[G2] $y_{zicjst} = \Omega_z X_{icjst} + \delta_{z1} \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \delta_{z2} \cdot (\varrho_2 \hat{\mu}_{behavior,jt}) + \tau_{st} + \delta_d \sum_{d=1}^{4} I_d + \nu_{zicjst}$.

The error term from [G2] can be written as $\nu_{zicjst} = \varepsilon_{zcjst} + \varepsilon_{zicjst}$. This includes a classroom-level shock (ε_{zcjst}) and a selection term (ε_{zicjst}). As such, there will be omitted variables or selection bias if either of the following two conditions does not hold:[3]

(1) $cov(\varepsilon_{zcjst}, \hat{\mu}_{zjt}) = 0$

(2) $cov(\varepsilon_{zicjst}, \hat{\mu}_{zjt}) = 0$

Because students may sort to teachers within schools, and teachers may sort to classrooms within schools, either of these conditions may be violated. Specifically, if some teachers within a school are systematically assigned to students who are above or below average in their school cohort in unobserved dimensions, it leads to $cov(\varepsilon_{zicjst}, \bar{\varepsilon}_{jz,i\neq t}) > 0$, so that condition (1) is violated and there is positive bias. Also, if certain teachers within a school are systematically

[3] The two conditions listed summarize the more detailed conditions; $cov(\varepsilon_{zcjst}, \theta_{zj}) = 0$, $cov(\varepsilon_{zcjst}, \bar{\varepsilon}_{zj,c\neq t}) = 0$, $cov(\varepsilon_{zcjst}, \bar{\varepsilon}_{jz,i\neq t}) = 0$, $cov(\varepsilon_{zicjst}, \theta_{zj}) = 0$, $cov(\varepsilon_{zicjst}, \bar{\varepsilon}_{zj,c\neq t}) = 0$, $cov(\varepsilon_{zicjst}, \bar{\varepsilon}_{zj,i\neq t}) = 0$.

assigned to better or worse classrooms in unaccounted-for dimensions (e.g. classrooms with better lights, classrooms in quieter areas of the school) then $cov(\varepsilon_{zcjst}, \bar{\varepsilon}_{zj,c\neq t}) > 0$, so that condition (1) is violated and there is positive bias. Note that both of these sources of bias involve the nonrandom allocation of students or classroom-level shocks *within* a school for a given cohort.

The across-cohort within-school identification strategy

One way to provide estimates that are not affected by the within-cohort biases outlined above is to exploit only the variation in estimated teacher quality across entire cohorts within a school (rather than across students or teachers within the same cohort or school). To introduce some notation, the variation in x that occurs *within* cohorts within a school is connoted by $\ddot{\Delta}x$ and the variation in x that occurs *across* cohorts within a school is connoted by $\bar{\Delta}x$.

Estimation of equation [12] (i.e. [G2] without school-by-year fixed effects) yields the error term $\nu_{zicjt} = \varepsilon_{zst} + \varepsilon_{zcjst} + \varepsilon_{zicjst}$, where ε_{zst} is a school-by-year time shock. By definition, the classroom-level shocks occur within schools, and the student selection occurs within schools so that $cov(\varepsilon_{zcjst}) = cov(\ddot{\Delta}\varepsilon_{zcjst})$ and $cov(\bar{\Delta}\varepsilon_{zcjst}) = 0$ and $cov(\varepsilon_{zicjst}) = cov(\ddot{\Delta}\varepsilon_{zicjst})$ and $cov(\bar{\Delta}\varepsilon_{zicjst}) = 0$. Similarly, by definition, all variation in the school-year shocks is across cohorts and none is within cohorts such that $cov(\varepsilon_{zst}) = cov(\bar{\Delta}\varepsilon_{zst})$ and $cov(\ddot{\Delta}\varepsilon_{zst}) = 0$. Consider, now, the average estimated teacher value-added for a given school *s* in a given year *t*, $\bar{\bar{\mu}}_{zjt}$=$[\bar{\theta}_{zj} + \bar{\bar{\varepsilon}}_{zj,c\neq t} + \bar{\bar{\varepsilon}}_{zj,i\neq t}]\lambda_{zj}$. Because there is no variation in $\bar{\bar{\mu}}_{zjt}$within a cohort by construction, $var(\ddot{\Delta}\bar{\bar{\mu}}_{zjt})$=0 , and therefore $var(\bar{\bar{\mu}}_{zjt}) = var(\bar{\Delta}\bar{\bar{\mu}}_{zjt})$. Consider using $\bar{\bar{\mu}}_{zjt}$ as an instrument for $\hat{\mu}_{zjt}$. In such an instrumental variables model, there is omitted variables or selection bias if $cov(\nu_{zicjst}, \hat{\mu}_{zjt}) \neq 0$. That is, there is omitted variables bias/selection bias if any one of the following three conditions does not hold:

(1) $cov(\varepsilon_{zcjst}, \bar{\bar{\mu}}_{zjt}) = cov(\ddot{\Delta}\varepsilon_{zcjst}, \bar{\Delta}\bar{\bar{\mu}}_{zjt}) = 0$ (*true by construction*)

(2) $cov(\varepsilon_{zicjst}, \bar{\bar{\mu}}_{zjt}) = cov(\ddot{\Delta}\varepsilon_{zicjst}, \bar{\Delta}\bar{\bar{\mu}}_{zjt}) = 0$ (*true by construction*)

(3) $cov(\varepsilon_{zst}, \bar{\bar{\mu}}_{zjt}) = cov(\bar{\Delta}\varepsilon_{zst}, \bar{\Delta}\bar{\theta}_{zj}) = 0$

By definition, because there is no within cohort variation in $\bar{\bar{\mu}}_{zjt}$, and all of the variation in ε_{zcjst} and ε_{zicjst} occur within a cohort, conditions (1) and (2) hold by construction. However, bias still exists if condition (3) does not hold. Condition (3) is that the arrival of a new teacher with high or low estimated value-added is unrelated to school-specific time shocks to outcomes. As such, the instrumental variables model that uses $\bar{\bar{\mu}}_{zjt}$ as an instrument for $\hat{\mu}_{zjt}$ is free from bias due to persistent classroom-level shocks and student selection to teachers within a school. However, the instrumental variables model may be biased if the timing of teacher mobility is not exogenous.

I implement the proposed instrumental variables strategy by estimating the following system of equations by two-stage least squares (2SLS).

[G3] $y_{zicjst} = \Omega_{0z}X_{icjst} + \delta_{0z1} \cdot (\varrho_1\hat{\mu}_{test,jt}) + \delta_{0z2} \cdot (\varrho_2\hat{\mu}_{behavior,jt}) + \kappa_{0d}\sum_{d=1}^{4} I_d + \tau_{0s} \cdot Year_t + \nu_{0zicjst}.$

[G4] $\hat{\mu}_{test,jt} = \Omega_{1z}X_{icjst} + \pi_{11}\bar{\bar{\mu}}_{test,jt} + \pi_{12}\bar{\bar{\mu}}_{behavior,jt} + \kappa_{1d}\sum_{d=1}^{4} I_d + \tau_{1s} \cdot Year_t + \nu_{1zicjst}.$

[G5] $\hat{\mu}_{behavior,jt} = \Omega_{2z}X_{icjst} + \pi_{21}\bar{\bar{\mu}}_{test,jt} + \pi_{22}\bar{\bar{\mu}}_{behavior,jt} + \kappa_{2d}\sum_{d=1}^{4} I_d + \tau_{2s} \cdot Year_t + \nu_{2zicjst}.$

To account for possible trends in outcomes at the school level, all models include school-specific linear time trends. Note that in equations [G3], [G4], and [G5] τ_{0s}, τ_{1s}, and τ_{2s} are linear time trends for school s, while τ_{st} is a school-year fixed effect. As an additional check on the exogeneity of the cohort-level changes, I also present the instrumental variables estimates on predicted outcomes (while excluding the covariates used to form the prediction, as in the test for selection on observables above). Note that, on average, each school-year cell has 9.17 teachers and the median number in each cell is 8.

Comparing patterns across the two models

Columns 1 and 2 in table G2 present the estimated OLS effects relying only on variation within school cohorts. The results are essentially unchanged from those in table 6, indicating that the main results presented were not confounded by school-level shocks that coincided with changes in teacher quality (i.e. value-added) across cohort within schools. Columns 3 and 4 present 2SLS regressions of high-school graduation and dropout outcomes based on estimated teacher value-added on ninth-grade skill measures. For both outcomes, results using the selection free variation across cohorts reveal that teachers that improve behaviors increase high-school graduation and reduce dropout outcomes. Even though the estimates are larger in the 2SLS models than the OLS models, (a) one cannot reject that the underlying effects are the same, and (b) the marginal effects in the 2SLS models are similar to the implied marginal effects of the conditional logit models.

Given that the 2SLS models are free from selection bias and also include school-specific linear time trends, the estimated relationships are likely to be real causal effects. However, as a final check on the 2SLS strategy, I estimate the 2SLS models on the predicted outcomes ($\tilde{y}_{zijt}$) while excluding seventh-grade behaviors and demographics from the main model. Results from this test are in columns 5 and 6. Consistent with no bias, there is no relationship between predicted outcomes and changes in teacher value-added across cohorts. Consistent with other studies that seek to validate teacher effects in value-added models (Chetty, Friedman, and Rockoff 2014b; Kane and Staiger 2008; Kane et al. 2013; and Bacher-Hicks, Kane, and Staiger 2015), I find little evidence of selection conditional on the rich set of covariates included in my models, and can rule out selection of student to teacher as the driver of the observed patterns.

TABLE G1

TESTING FOR SELECTION ON OBSERVABLES USING A LIMITED SET OF CONTROL VARIABLES

	1	2	3	4	5	6	7	8
	Graduate from High School		Drop Out of School		Predicted: Graduate from High School		Predicted: Drop Out of School	
Teacher Value-Added: 9th Grade Test Score	0.00144*	0.00152**	-0.000324	-0.000367	-0.000273+	-0.000289+	-0.000101	-0.000129
	[0.000571]	[0.000581]	[0.000295]	[0.000297]	[0.000160]	[0.000160]	[0.000156]	[0.000156]
Teacher Value-Added: 9th Grade Behaviors	0.0134**		-0.00388*		9.87E-06		-0.000767	
	[0.00324]		[0.00195]		[0.000753]		[0.000776]	
Teacher Value-Added: 10th Grade GPA		0.0154**		-0.00314		0.00104		0.00063
		[0.00582]		[0.00309]		[0.00143]		[0.00141]
Observations	896,956	896,956	896,956	896,956	896,956	896,956	896,956	896,956

Note: Robust standard errors in brackets are adjusted for two-way clustering at the teacher and student levels.
The models include track fixed effects and year fixed effects, incoming outcomes in eighth grade (math and reading scores in eighth grade, repeater status in eighth grade, ever suspended in eighth grade, GPA in eighth grade (for high-school courses only), and attendance in eighth grade), classroom averages of these lagged outcomes, and the number of honors courses taken during ninth grade. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.
Predicted outcomes are based on a logistic regression of each outcome on seventh-grade math and reading scores, seventh-grade repetition, suspensions in seventh grade, absences in seventh grade, parental education, gender, and ethnicity. In columns 5 through 8, which use predicted outcomes as the dependent variable, the seventh-grade outcomes, demographic variables, and classroom means of these variables are excluded as controls.
** p<0.01, * p<0.05, + p<0.1

TABLE G2
2SLS REGRESSIONS USING COHORT-LEVEL VARIATION IN TEACHER QUALITY

	1	2	3	4	5	6
	OLS with School-Year Fixed Effects		2SLS using Average Teacher Quality in the School-Cohort as an Instrument [a]			
	Graduate	Dropout	Graduate	Dropout	Predicted Graduate [b]	Predicted Dropout [b]
Teacher Value-Added: 9th Grade Test	0.000752*	-0.000205	0.00362	-0.00161	0.000579	0.000688
	[0.000471]	[0.000265]	[0.00230]	[0.00115]	[0.00114]	[0.00112]
Teacher Value-Added: 9th Grade	0.01287**	-0.00320+	0.0215+	-0.0149*	-0.000386	-0.00403
	[0.002794]	[0.001795]	[0.0124]	[0.00722]	[0.00566]	[0.00550]
Teacher Value-Added: 9th Grade Test	0.000858+	-0.000240	0.00316	-0.00153	0.000325	0.000438
	[0.000481]	[0.000269]	[0.00232]	[0.00116]	[0.00114]	[0.00112]
Teacher Value-Added: 10th Grade GPA	0.01252**	-0.00255	0.0558**	-0.0241*	0.0139	0.00861
	[0.00482]	[0.00269]	[0.0208]	[0.0107]	[0.0108]	[0.0104]
School-Track Fixed Effects	N	N	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y	Y	Y
School-Year Fixed Effects	Y	Y	-	-	-	-
School-Specific Linear Time Trends	-	-	Y	Y	Y	Y
All controls	Y	Y	Y	Y	N	N
Observations	896956	896956	891844	891844	891844	891844

Note: Robust standard errors in brackets are adjusted for clustering at both the teacher and student level.
The models in columns 1 through 4 include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models in columns 1 through 4 also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.
a. The excluded instruments in the 2SLS models are the average estimated teacher effects at the school-year level. The first stage F-statistics are greater than 1000 in all models.
b. Predicted outcomes are based on a logistic regression of each outcome on 7th grade math and reading scores, 7th grade repetition, suspensions in 7th grade and absences in 7th grade, parental education, gender, and ethnicity. In columns 5 and 6 that use predicted outcomes as the dependent variable, the 7th grade outcomes, the demographic variables, and the classroom means of these variables are excluded as controls.
** p<0.01, * p<0.05, + p<0.1

Testing for bias due to the confounding effect of other teachers

The second identifying assumption is that, conditional on the track variables and controls, the quality (i.e. value-added) of a student's teacher in one subject is uninformative about the quality of their other subject teachers. I test the validity of this identifying assumption in three ways. First, for each student I merge the estimated English teacher value-added into the math student data. This results in a dataset of all students who have both estimated math teacher value-added and English teacher value-added on both test scores and behaviors. If there were sorting of teachers to groups of students *within tracks* such that the estimated teacher value-added did not isolate the effects of the individual teacher, but reflected the contribution of a different teacher, the math and English teacher value-added estimates would be positively correlated.

Table G3 presents the correlations between teacher value-added on the different skill measures across the two subject teachers (math and English). The correlations between value-added across the two subject teachers are close to zero (note that some are negative and some are positive). This suggests that, conditional on controls, the quality of the math teacher is unrelated to the quality of the English teacher. This also is compelling evidence that there is not a third teacher (in a subject other than math or English) who is driving the effects. If there were such a third teacher driving the effects, then the same teacher that leads to a spurious positive math teacher value-added will lead to a spurious positive English teacher value-added – generating a spurious positive correlation between the two. However, this is clearly not the case empirically.

As an additional test of this assumption, I use the same data (as described above) and regress math teacher value-added on the estimated English teacher value-added, conditional on all the controls. In such models (presented in table G4), for both test scores and behaviors, one fails to reject the null hypothesis that the estimated English teacher value-added is unrelated to the estimated math teacher value-added at the 10 percent level. Finally, to show that the estimated effects are not driven by the contributions of other subject teachers, I estimate all the main models while including indicator variables for the other subject teachers. Table G5 presents the main results where I include indicator variables for each math teacher when the own teacher is the English teacher, and include indicator variables for each English teacher when the own teacher is the math teacher (that is, the models include fixed effects for the *other*-subject teacher). All such models cluster standard errors at both the math teacher and the English teacher levels. The main results are robust to including other teacher fixed effects.

In sum, all of the empirical tests suggest that conditional on the controls for tracking and sorting, the quality of a student's teacher in one subject is unrelated the quality of that student's teachers in other subjects.

TABLE G3

CORRELATIONS BETWEEN ENGLISH AND MATH TEACHER VALUE-ADDED

	Math Teacher: Test-score Value-Added	Math Teacher: Behaviors Value-Added	Math Teacher: 10^{th} Grade GPA Value-Added	English Teacher: Test-score Value-Added	English Teacher: Behaviors Value-Added	English Teacher: 10^{th} Grade GPA Value-Added
Math Teacher: Test-score Value-Added	1					
Math Teacher: Behaviors Value-Added	0.2582	1				
Math Teacher: 10^{th} Grade GPA Value-Added	0.2144	0.3391	1			
English Teacher: Test-score Value-Added	0.0088	-0.0018	0.0022	1		
English Teacher: Behaviors Value-Added	0.0102	0.0078	-0.0064	0.1292	1	
English Teacher: 10^{th} Grade GPA Value-Added	-0.0032	0.0056	0.0087	0.1093	0.2067	1

TABLE G4

RELATIONSHIP BETWEEN ENGLISH AND MATH TEACHER VALUE-ADDED WITHIN TRACKS

	1	2	3	4	5	6
	Math Teacher: Test-score Value-Added	Math Teacher: Behaviors Value-Added	Math Teacher: 10^{th} Grade GPA Value-Added	Math Teacher: Test-score Value-Added	Math Teacher: Behaviors Value-Added	Math Teacher: 10^{th} Grade GPA Value-Added
English Teacher: Test-score Value-Added	0.00572			0.00545		
	[0.00457]			[0.00457]		
English Teacher: Behaviors Value-Added		-0.000266			-0.000261	
		[0.00100]			[0.00100]	
English Teacher: 10^{th} Grade GPA Value-Added			-0.00298			-0.00286
			[0.00288]			[0.00289]
Year Effects	Y	Y	Y	Y	Y	Y
School-Track Effects	Y	Y	Y	Y	Y	Y
Controls	N	N	N	Y	Y	Y
Observations	348,514	348,514	348,514	346,223	346,223	346,223

Robust standard errors in brackets

These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed

** p<0.01, * p<0.05, + p<0.1

TABLE G5
ROBUSTNESS TO INCLUDING TEACHER EFFECTS IN THE OTHER SUBJECT

	Graduate				Dropout			
	1	2	3	4	5	6	7	8
Teacher: Test-score Value-Added	0.00118*	0.000752	0.00111+	0.00104+	-0.000315	-0.000205	-7.33E-05	6.09E-05
	[0.000546]	[0.000471]	[0.000619]	[0.000620]	[0.000290]	[0.000267]	[0.000310]	[0.000317]
Teacher: Behaviors Value-Added	0.0146**	0.0129**	0.0121**	0.0136**	-0.00407*	-0.00335+	-0.00371+	-0.00369
	[0.00319]	[0.00288]	[0.00369]	[0.00320]	[0.00192]	[0.00178]	[0.00190]	[0.0112]
	9	10	11	12	13	14	15	16
Teacher: Test-score Value-Added	0.00130*	0.000858+	0.00114+	0.00107+	-0.000363	-0.00024	-0.000119	3.08E-05
	[0.000556]	[0.000483]	[0.000624]	[0.000628]	[0.000292]	[0.000268]	[0.000310]	[0.000315]
Teacher: 10^{th} Grade GPA Value-Added	0.0146**	0.0125**	0.0152*	0.0182	-0.00307	-0.00254	-0.00237	-0.00341
	[0.00565]	[0.00473]	[0.00644]	[0.0153]	[0.00305]	[0.00266]	[0.00325]	[0.00933]
Track-School Effects	Y	Y	Y	Y	Y	Y	Y	Y
School Year Effects	N	Y	N	Y	N	Y	N	Y
Other Teacher Fixed Effect	N	N	Y	Y	N	N	Y	Y
Observations	891,726	891,726	891,726	891,726	891,726	891,726	891,726	891,726

Note: Robust standard errors in brackets adjusted for two-way clustering at both the math and English teacher levels.
These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed
** p<0.01, * p<0.05, + p<0.1

Appendix H

Effects by Subject

The results thus far have analyzed English and math teachers together. I relax this restriction and show effects for English and math teachers separately. This is accomplished by interacting the estimated teacher value-added with indicators for the subject and including these interactions in the regression model. Specifically, in [H1], I estimate the following where $Math_j$ is an indicator variable equal to 1 if teacher j is a math teacher (i.e. the subject is algebra I, geometry, or algebra II), and $English_j$ is an indicator variable equal to 1 if teacher j is an English teacher.

[H1] $y_{zicjst} = \Omega_z X_{icjst} + \delta_{z1,math} \cdot (\varrho_1 \hat{\mu}_{test,jt}) \cdot Math_j + \delta_{z1,English} \cdot (\varrho_1 \hat{\mu}_{test,jt}) \cdot English_j + \delta_{z2,Math} \cdot (\varrho_2 \hat{\mu}_{behavoir,jt}) \cdot Math_j + \delta_{z2,english} \cdot (\varrho_2 \hat{\mu}_{behavoir,jt}) \cdot English_j + \delta_d \sum_{d=1}^{4} I_d + \nu_{zicjst}.$

The coefficient estimates of $\delta_{z1,math}$ and $\delta_{z1,English}$ represent the marginal effect of increasing the math teacher and the English teacher value-added on test scores by one standard deviation, respectively. The coefficient estimates of $\delta_{z2,math}$ and $\delta_{z2,English}$ represent the marginal effect of increasing the math teacher and the English teacher value-added on behaviors by one standard deviation, respectively.

The estimates are presented in tables H1 and H2. Table H1, column 1 shows the estimated effect on test scores. As expected, teacher value-added on test-scores predict test scores, and the effects are larger for math teachers (0.1σ) than for English teachers (0.033σ). While both test-score value-added estimates have statistically significant effects on test scores at the 1 percent level, behaviors value-added has no effect on test scores in either subject. Column 2 presents effects on behaviors. The results indicate that the marginal effect of increasing teacher behaviors value-added by one standard deviation on behaviors is somewhat larger in math than in English. Indeed, one can reject that the two marginal effects are equal at the 5 percent level. Column 3 presents effects on whether a student is enrolled in tenth grade. Similar to the effects on behaviors, the results indicate that the marginal effect of increasing teacher behaviors value-added by one standard deviation on tenth-grade enrollment is larger in math than in English. However, looking at tenth-grade GPA, high-school graduation, and dropout, the marginal effects of increasing teacher behaviors value-added are larger for English teachers than for math teachers. This pattern of larger effects for English teachers is also present for GPA in 12th grade. However, the marginal effect is larger for math teachers in predicting whether a student takes the SAT and college intentions. Despite some differences in point estimates, many of the differences across the subjects are not statistically significant at traditional levels. In sum, there is little evidence that the effects vary systematically across the two subjects.

TABLE H1

EFFECTS OF TEACHERS ON SKILL MEASURES AND THEIR EFFECTS ON VARIOUS LONGER-RUN OUTCOMES BY SUBJECT

	1	2	3	4	5	6
	Test-score in 9th Grade	Behaviors in 9th Grade	Enrolled in 10th Grade	10th Grade GPA	Dropout of School	Graduate High School
English Teacher Value-Added: 9th Grade Test Score	0.0327**	-0.00179	-0.000171	0.00109	-0.00017	0.00118
	[0.00325]	[0.00236]	[0.000766]	[0.00182]	[0.000442]	[0.000780]
Math Teacher Value-Added: 9th Grade Test Score	0.101**	0.0118**	0.00137*	0.00655**	-0.000522	0.00137+
	[0.00379]	[0.00210]	[0.000659]	[0.00162]	[0.000391]	[0.000762]
English Teacher Value-Added: 9th Grade Behaviors	-0.00901	0.0492**	0.0187**	0.0144+	-0.00452*	0.0156**
	[0.0112]	[0.0111]	[0.00334]	[0.00828]	[0.00202]	[0.00334]
Math Teacher Value-Added: 9th Grade Behaviors	-0.0633	0.137**	0.0448**	0.00555	0.00129	0.00169
	[0.0459]	[0.0384]	[0.0107]	[0.0238]	[0.00546]	[0.0110]
pr(Test-Score Value-Added Same)	0.00	0.00	0.13	0.02	0.55	0.86
pr(Behavior Value-Added Same)	0.25	0.03	0.02	0.73	0.32	0.23
	7	8	9	10	11	12
English Teacher Value-Added: 9th Grade Test Score	0.0327**	-0.000347	0.000326	0.00115	-0.000248	0.00144+
	[0.00320]	[0.00251]	[0.000817]	[0.00178]	[0.000444]	[0.000798]
Math Teacher Value-Added: 9th Grade Test Score	0.0998**	0.0125**	0.00187**	0.00623**	-0.000502	0.00122
	[0.00377]	[0.00215]	[0.000680]	[0.00159]	[0.000386]	[0.000767]
English Teacher Value-Added: 10th grade GPA	-0.022	0.00896	0.00761	0.0329+	-0.00478	0.0171*
	[0.0241]	[0.0200]	[0.00630]	[0.0195]	[0.00395]	[0.00725]
Math Teacher Value-Added: 10th grade GPA	0.0288	0.0969**	0.0134+	0.0235	0.000345	0.00982
	[0.0369]	[0.0254]	[0.00769]	[0.0228]	[0.00447]	[0.00882]
pr(Test-Score Value-Added Same)	0.000	0.000	0.149	0.0334	0.667	0.846
pr(Behavior Value-Added Same)	0.248	0.006	0.56	0.753	0.385	0.522
Observations	942,291	941,855	942,291	728,529	891,868	891,868

Note: Standard errors in brackets are adjusted for two-way clustering at the teacher and student level.

These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.

** p<0.01, * p<0.05, + p<0.1

TABLE H2

EFFECTS OF TEACHERS ON SKILL MEASURES AND THEIR EFFECTS ON VARIOUS LONGER-RUN OUTCOMES BY SUBJECT CONT'D

	1	2	3	4	5	6
	Take the SAT	Intend to Attend 4-year College	GPA in 12th grade	SAT: Math Score	SAT: Verbal Score	SAT: Writing Score
English Teacher Value-Added: 9th Grade Test Score	-8.93E-05	0.00152	-0.00217+	0.00103	0.0376	0.708**
	[0.000964]	[0.00110]	[0.00127]	[0.169]	[0.189]	[0.197]
Math Teacher Value-Added: 9th Grade Test Score	0.00201*	0.000731	0.00428**	0.442**	-0.168	0.148
	[0.000998]	[0.00114]	[0.00120]	[0.144]	[0.147]	[0.167]
English Teacher Value-Added: 9th Grade Behaviors	0.0106**	0.00981*	0.0243**	-0.562	-0.276	0.207
	[0.00392]	[0.00462]	[0.00598]	[0.776]	[0.749]	[0.782]
Math Teacher Value-Added: 9th Grade Behaviors	0.0280*	0.01585	-0.00324	1.669	2.277	-0.0139
	[0.0141]	[0.01518]	[0.0163]	[2.262]	[2.091]	[2.367]
pr(Test-score Value-Added Same)	0.13	0.37	0.00	0.07	0.41	0.02
pr(Behavior Value-Added Same)	0.24	0.68	0.11	0.31	0.38	0.82
	7	8	9	10	11	12
English Teacher Value-Added: 9th Grade Test Score	0.000144	0.00168	-0.00174	-0.0349	0.0428	0.737**
	[0.000978]	[0.00109]	[0.00128]	[0.167]	[0.189]	[0.198]
Math Teacher Value-Added: 9th Grade Test Score	0.00232*	0.000801	0.00409**	0.452**	-0.121	0.132
	[0.000988]	[0.00114]	[0.00119]	[0.142]	[0.144]	[0.164]
English Teacher Value-Added: 10th grade GPA	0.00702	0.01115	0.0246*	1.421	-0.957	-1.63
	[0.00845]	[0.01113]	[0.0116]	[1.587]	[1.554]	[1.716]
Math Teacher Value-Added: 10th grade GPA	0.00953	0.012214	0.00754	1.069	-0.69	1
	[0.0124]	[0.01317]	[0.0142]	[1.882]	[1.800]	[2.119]
pr(Test-score Value-Added Same)	0.117	0.346	0.001	0.026	0.485	0.018
pr(Behavior Value-Added Same)	0.866	0.563	0.351	0.887	0.909	0.333
Observations	789627	789627	701813	401744	401744	401744

Note: Standard errors in brackets are adjusted for two-way clustering at the teacher and student level.

These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed

** p<0.01, * p<0.05, + p<0.1

Appendix I

TABLE I1
OBSERVABLE TEACHER CORRELATES OF THE BEHAVIOR INDEX

	1	2	3	4	5	6	7	8
	Without Teacher Fixed Effects				With Teacher Fixed Effects			
	Test Scores	Behavior index	Graduate	Dropout	Test Scores	Behavior index	Graduate	Dropout
Racial Match	0.00542	0.00261	0.00498**	0.000547	0.00301	0.000704	0.00655**	0.000198
	[0.00390]	[0.00429]	[0.00179]	[0.000891]	[0.00305]	[0.00463]	[0.00207]	[0.00102]
Gender Match	0.0472**	0.00485	3.10E-06	-0.00022	0.0472**	0.00663	0.000761	-0.000333
	[0.00519]	[0.00421]	[0.00198]	[0.00101]	[0.00523]	[0.00420]	[0.00199]	[0.00101]
Ln(Years of Experience)	0.00106	-0.000858	-0.000719	0.000343	0.0178**	-0.000352	-0.00236	-0.00114
	[0.00298]	[0.00196]	[0.000675]	[0.000371]	[0.00685]	[0.00785]	[0.00274]	[0.00150]
Certified	0.0151+	0.00309	0.00275	-0.00064	0.00921	-0.0017	0.00392	0.000582
	[0.00781]	[0.00620]	[0.00208]	[0.00117]	[0.00888]	[0.0103]	[0.00351]	[0.00191]
Average Test Score	0.00148	-0.00271	-4.13E-05	0.000211	0.0861*	0.0254	-0.00779	-0.00169
	[0.00321]	[0.00187]	[0.000629]	[0.000330]	[0.0407]	[0.0420]	[0.0175]	[0.00914]
Advanced Degree	-0.0017	0.0029	0.00135	-3.35E-05	0.00285	0.00372	0.00791+	0.000314
	[0.00479]	[0.00287]	[0.000993]	[0.000511]	[0.00996]	[0.0107]	[0.00471]	[0.00233]
75^{th}%ile SAT at College	9.56e-05*	-4.24e-05+	5.97E-06	-2.08E-06	0.00458**	-0.00113	0.000699*	9.41E-05
	[4.29e-05]	[2.52e-05]	[8.41e-06]	[4.54e-06]	[0.00106]	[0.00133]	[0.000331]	[0.000227]
Fully Licensed	0.0230**	-0.00361	0.0022	-0.00115	0.0198*	-0.0102	0.00426	-0.00196
	[0.00641]	[0.00498]	[0.00180]	[0.000951]	[0.00811]	[0.00988]	[0.00328]	[0.00183]
Licensed in Math	0.0415**	-0.00426	-0.00462	0.00166	0.0476	-0.0164	0.00985	-0.0012
	[0.0148]	[0.0101]	[0.00461]	[0.00221]	[0.0295]	[0.0248]	[0.00927]	[0.00433]
Observations	726,694	726,694	726,694	726,694	726,694	726,694	726,694	726,694

Standard errors in brackets are adjusted for two-way clustering at the teacher and student level.
These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed
** p<0.01, * p<0.05, + p<0.1

Appendix J

Testing Additivity of Teacher value-Added in the Production of Student Skills

The model presented in section II makes some important functional form assumptions regarding the production of student skills. One key implication of the model is that the value-added of individual teachers are additively separable. To explore whether the value-added of teachers in producing student skills are additive across subjects, I link each student to the estimated value-added of both their math and English teachers and do not estimate the long run effects by teacher subject separately (that is, I have one observation per student). Because English teachers are solely responsible for English scores and math teachers are solely responsible for math scores, I focus on the production of behaviors. Note that the estimated coefficient on the math teacher value-added on English scores is 0.00134 (p-value = 0.24) and the estimated coefficient on the English teacher value-added on math scores is 0.00249 (p-value = 0.338).

Table J1 presents the estimated effects of the English teacher on behaviors, the estimated effects of the math teacher on behaviors, and the interaction between the two. Under the additive model, the interaction between the value-added of the two subject teachers will be zero. Column 2 shows that while each teacher independently impacts behaviors, the interactions between the two teachers' value-added does not. Column 5 shows the same basic pattern in predicting tenth-grade GPA. As an additional check I explore whether the interaction predicts high school graduation in columns 3 and 6. In neither case are the interactions statistically significant. I also explore whether the interaction of the teacher value-added on test scores across subjects predicts high school graduation (not shown), and the coefficient of the interaction is 0.00078 (p-value = 0.518). In sum, one cannot reject that the impacts of teacher value-added across subjects on skills (and long-term outcomes) are additive.

TABLE J1
TESTING FOR INTERACTIONS BETWEEN TEACHERS ACROSS SUBJECTS

	1	2	3	4	5	6
	Behaviors in 9th Grade		Graduate	GPA in 10th Grade		Graduate
Math Teacher Value-Added: Behaviors	0.0627+	0.0618	0.00399			
	[0.0343]	[0.0402]	[0.0134]			
English Teacher Value-Added: Behaviors	0.0485**	0.0485**	0.0130**			
	[0.0125]	[0.0124]	[0.00378]			
Math*English Teacher Value-Added Interaction: Behaviors		-0.0632	0.0299			
		[0.183]	[0.0656]			
Math Teacher Value-Added: 10th Grade GPA				0.0394	0.0422+	0.00711
				[0.0240]	[0.0240]	[0.0107]
English Teacher Value-Added: 10th Grade GPA				0.0392+	0.0390+	0.0173*
				[0.0230]	[0.0230]	[0.00835]
Math*English Teacher Value-Added Interaction: 10th Grade GPA					0.17	-0.068
					[0.128]	[0.0582]
Observations	369,547	369,547	360,364	306,419	306,419	360,364
F-test(Subject effects are the same)	0.758	0.773	0.517	0.996	0.927	0.45
F-test(interaction=0)		0.73	0.649		0.183	0.242

Note: Robust standard errors in brackets adjusted for two-way clustering at both the English teacher and math teacher levels.
These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged outcomes (math scores, reading scores, repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors, both eighth-grade and seventh-grade test scores, and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed

** p<0.01, * p<0.05, + p<0.1.

Appendix K

Robustness to Excluding Lagged Test-Score Controls

Given that the controls for lagged GPA are imperfect, it is helpful to assess how robust the main findings are to the exclusion of controls for test scores. To assess this, I estimate teacher value-added on behaviors and test scores excluding all test score variables in eighth grade and seventh grade (also excluding the classroom-level means of these test score variables). I then compute the covariance of teacher value-added as before, and rescale the estimated out-of-sample teacher value-added estimates so that the coefficients of the leave-year-out estimates is equal to the impact of increasing teacher value-added by one standard deviation (from the naïve model with no test score controls). Table K1 presents the estimated effects on high school graduation and dropping out.

Even though the model excludes eighth-grade and seventh-grade test scores (and their classroom averages), the pattern of results is very similar to that in table 6. As expected, test-score value-added is slightly more predictive of the long-term outcomes, but not greatly so. In fact, the standard errors are sufficiently large that one cannot reject the null hypothesis that the estimated effects on high-school graduation and dropping out are the same in models that account for lagged test scores as in those that do not. Importantly, for both longer-term outcomes, even when lagged test score controls are not included, including teacher value-added on ninth-grade behaviors increases the variance of the explained teacher impacts by over 200 percent. This shows that (a) the controls for tracking are sufficient to account for much potential bias due to sorting, and (b) the basic patterns documented in this paper are quite robust.

TABLE K1

EFFECTS OF TEACHER VALUE-ADDED ON HIGH SCHOOL COMPLETION: WITHOUT TEST-SCORE CONTROLS

	1	2	3	4	5	6
	Graduate High School			Dropout of high School		
	Linear Probability Model			Linear Probability Model		
Teacher Value-Added: 9th Grade Test Score	0.00213*	0.00190*	0.00189*	-0.000752	-0.000637	-0.000752
	[0.000931]	[0.000928]	[0.000935]	[0.000464]	[0.000465]	[0.000468]
Teacher Value-Added: 9th Grade Behaviors		0.0108**			-0.00533*	
		[0.00370]			[0.00225]	
Teacher Value-Added: 10th Grade GPA			0.00822*			0.00121
			[0.00383]			[0.00186]
% Increase in explained variance		190%	85%		367%	1%
School-Track Effects	Y	Y	Y	Y	Y	Y
Year Effects	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	891,868	891,868	891,868	891,868	891,868	891,868

Note: Robust standard errors in brackets adjusted for two-way clustering at the teacher and student levels.
These regressions are based on the pooled sample across both math and English teachers. In total, there are 11,857 teachers across the two subjects. All models include track fixed effects and year fixed effects, the number of honors courses taken during ninth grade, student-level demographics (parental education, ethnicity, and gender), lagged behaviors (repeater status, suspensions, and attendance all in both seventh and eighth grades, and GPA in eighth grade [for high-school courses only]). Models also include classroom averages of eighth-grade behaviors and student demographics. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.
Note: Unlike in Table 6, the teacher effects in this model are estimated without controls for eighth-grade test score, seventh-grade test scores, or classroom averages of these test scores controls.
To compute the increase in the variance explained, I compute the variance of the fitted values for each teacher in models without behaviors value-added (i.e. $a = var(\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}))$, and in models with value-added on both (i.e. $b = var[\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}) + \hat{\delta}_2 \cdot (\varrho_2 \hat{\mu}_{behavior,jt})]$. The percentage increase in explained variability from also including behaviors value-added (versus test-score value-added alone) is $100 \times ((\mathrm{b} \div \mathrm{a}) - 1)$.
** p<0.01, * p<0.05, + p<0.1