Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-16-03

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

C. Kirabo Jackson
Associate Professor of Human Development and Social Policy
Faculty Fellow, Institute for Policy Research
Northwestern University

Version: March 6, 2016

DRAFT

Please do not quote or distribute without permission.

ABSTRACT

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 skill. Results show that teachers have effects on skills not measured by test-scores, but reflected in absences, suspensions, course grades, and on-time grade progression. Teacher effects on these non-test-score outcomes in 9th grade predict effects on high-school completion and predictors of college-going—above and beyond their effects on test scores. Relative to using only test-score measures of teacher quality, including both test-score and non-test-score measures more than doubles the predictable variability of teacher effects on these longer-run outcomes.

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

C. Kirabo Jackson, 6 March, 2016
Northwestern University and NBER

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 skill. Results show that teachers have effects on skills not measured by test-scores, but reflected in absences, suspensions, course grades, and on-time grade progression. Teacher effects on these non-test-score outcomes in 9th grade predict effects on high-school completion and predictors of college-going—above and beyond their effects on test scores. Relative to using only test-score measures of teacher quality, including both test-score and non-test-score measures more than doubles the predictable variability of teacher effects on these longer-run outcomes. (JEL I21, J00)

There is widespread agreement that teachers are a key component of the schooling environment. At the broadest level, a quality teacher is one that teaches students the skills needed to be productive adults (Douglass 1958; Jackson et. al. 2014). However, economists have focused on test-score measures of teacher quality (called value added) because they are often the best available measure of student skills.[2] In an influential paper, Chetty, Friedman, and Rockoff (2014b) show that teachers who improve test scores (i.e. high value added teachers) improve students' longer run outcomes such as high school completion, college-going, and earnings. However, 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.[3] This literature provides reason to suspect that teachers may impact skills that go undetected by test scores, but are nonetheless important for students' long run success. Because 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.

To speak to these issues, this paper explores the extent to which teacher effects on measures of noncognitive skills predict effects on longer-run outcomes that go undetected by test score

[1] I thank David Figlio, Jon Guryan, Simone Ispa-Landa, Clement Jackson, Mike Lovenheim, James Pustejovsky, Jonah Rockoff, Alexey Makarin, and Dave Deming for insightful comments. I also thank Kara Bonneau from the NCERDC and Shayna Silverstein. This research was supported by funding from the Smith Richardson Foundation.
[2] Having a teacher at the 85th versus the 15th percentile of the test score value-added distribution is found to increase test score by between 8 and 20 percentile points (Kane and Staiger, 2008; Rivkin, Hanushek, and Kain, 2005).
[3] See Lindqvist and Vestman, 2011; Heckman and Rubinstein, 2001; Waddell, 2006; 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 (Fredricksson et al 2012; Heckman, Pinto, and Savelyev 2013).

effects.[4] This paper (a) extends the standard value-added model to estimate teacher effects on both test scores and also proxies for noncognitive skills, (b) documents the extent to which teachers who raise tests scores also raise proxies for noncognitive skills and *vice versa*, and (c) documents the extent to which a teacher's estimated effects on proxies for noncognitive skills predict effects on longer-run outcomes above and beyond that predicted using test score value-added alone.

This project employs rich administrative data on all public school 9th graders in North Carolina from 2005 to 2012. These data contain student scores on Algebra I and English I exams in 9th grade linked to their subject teachers. To obtain measures of student skills in 9th grade that *may* not be well-captured by test scores, I follow a large literature that uses behavioral outcomes as proxies for noncognitive skills (e.g. Heckman, Stixrud, and Urzua 2006, Lleras 2008, Bertrand and Pan 2013, Kautz 2014).[5] The outcomes used are suspensions, attendance, course grades, and on-time grade progression; each of which has been shown to be sensitive to well-known measures of noncognitive skills developed by psychologists. To summarize the behavioral outcomes with a single variable and to reduce measurement error, I compute an underlying factor (i.e. a weighted average of absences, suspensions, grades, and grade progression) that explains covariance across these outcomes. I refer to this weighted average of 9th grade behaviors as the behavioral factor. I am able to examine effects on longer-run student outcomes such as high-school completion, SAT-taking, and intentions to attend college that are collected through 12th grade. Even though these longer-run outcomes are measured at a young age, they include strong predictors of college going, and high-school dropout is a strong predictor of crime, employment, and earnings. Accordingly, these outcomes are economically important and worthy of study in their own right.

To motivate the empirical work, I extend the standard value-added model that assumes that ability is unidimensional (Todd and Wolpin 2003). In the extended model, student outcomes are a function of their stock of both cognitive and noncognitive dimensions of skill (Heckman, Stixrud, and Urzua 2006). The model demonstrates that as long as test scores and behavioral outcomes do not reflect the same exact mix of student skills then (a) there may be teachers who improve long-run outcomes that do not raise test scores, and (b) one can better predict a teacher's effect on long-

[4] Alexander, Entwisle, and Thompson (1987), Ehrenberg, Goldhaber, and Brewer (1995), Downey and Shana (2004), Jennings & 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] The basic idea is intuitive. One can infer that a student who acts out, skips class, and does not hand in homework likely has lower motivation and weaker interpersonal skill than a student who does not in exactly the same way one infers that a student who scores higher on tests likely has higher cognitive skill than a student who does not.

run student outcomes using effects on both test scores and behavioral outcomes in 9th grade.

This paper uses value-added models to identify teacher effects on test scores and on proxies for noncognitive skill. Teacher effects from value-added models have been validated in many settings (i.e. Kane and Staiger 2008; Kane, McCaffrey, Miller and Staiger 2013; Chetty, Friedman, and Rockoff 2014a; Bacher-Hicks, Kane, and Staiger 2015). However, to ensure that the teacher effect estimates presented in this paper can be interpreted causally, all models include a rich set of covariates, and I present several empirical tests to show that the effects are not biased. Using these value-added models, 9th grade teachers have meaningful effects on both test scores and the behavioral outcomes. Interestingly, teacher effects on test scores and the behavioral factor are weakly correlated (ρ=0.16), and teachers that systematically raise one outcome (test scores or behaviors) have virtually no effect on the other outcome. These patterns suggest that value-added and effects on behaviors (i.e. proxies for noncognitive skills) measure changes on distinct skills.

To explore whether teacher effects on the behavioral factor predict effects on longer-run outcomes above and beyond test score value-added, I link the 9th grade student data and the estimated teacher effects to data on high-school dropout, high-school graduation, SAT taking, and stated intentions to attend college. In models that predict high-school graduation using test score value-added only, a one standard deviation increase in value added raises the likelihood of high-school graduation by 0.13 percentage points. However, when also including teacher effects on the behavioral factor, a one standard deviation increase in value added leads to 0.11 percentage points higher likelihood of graduation, and a one standard deviation increase in the teacher's behavioral factor effect leads to 0.78 percentage points higher likelihood of graduating high school. These effect sizes are on the same order of magnitude as the college-going effects presented in Chetty et al (2014b). Including both effects more than doubles the predictable teacher-level variability in high-school graduation. Patterns are similar for dropout, SAT taking, and college plans.

This study demonstrates that non-test-score outcomes can identify teachers who improve longer-run outcomes but have no effect on test scores. The results support an idea that many believe to be true but had not previously been shown – that teacher effects on test scores capture only a fraction of their effect on human capital. This underscores the need for holistic evaluation approaches that account for effects on both cognitive and noncognitive skill. Because the non-test-score outcomes used (i.e. course grades, and suspension) can be manipulated by teachers, using them directly for accountability or evaluation purposes is unwise. However, I present some feasible

policy uses. The study also has implications for the broader literature. First, the patterns provide an explanation for why Chamberlain (2013) finds that value-added estimates may reflect less than one-fifth of the total effect of teachers. Also, the importance of teacher effects on skills not well measured by test scores offers an explanation for why teacher effects tend to fade over time (Jacob, Lefgren, and Sims 2010) despite teachers having meaningful effects on students in the long run.

The remainder of this paper is organized as follows: Section II presents the theoretical framework. Section III describes the data. 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 discuss possible uses for policy. Section VII concludes.

II Theoretical Framework

The standard value-added model assumes that student ability is one-dimensional (Todd and Wolpin 2003). I extend this model such that student outcomes are functions of *both* cognitive and noncogntive skills (Heckman, Stixrud, and Urzua 2006).[6] This extension allows for the possibilty that teachers can improve a set of skills that lead to improved longer-run outcomes but are not relfected in improved test scores. I derive some key testable implications from the model.

II.1 Model Setup

Student Skill: Prior to 9th grade, each student *i* has a stock of cognitive and noncognitive skill described by vector $\nu_i = (\nu_{c,i}, \nu_{n,i})$, where the subscripts *c* and *n* denote the cognitive and noncognitive dimensions, respectively. This stock reflects an initial endowment and the cumulative effect of all school and parental inputs on students' incoming skills (Todd and Wolpin 2003). Each 9th grade teacher *j* has a mean-zero vector $\omega_j = (\omega_{c,j}, \omega_{n,j})$ that describes teacher *j's* "value added" to each of the two dimensions of student skill during 9th grade. At the end of 9th grade, student *i* exposed to teacher *j* has total ability vector $\alpha_{ij} = \upsilon_i + \omega_j$.[7]

Outcomes: There are multiple short-run outcomes y_s for each student *i* measured at the end of 9th grade. Each 9th grade outcome y_s is a function of the two-dimensional skill vector given by [1], where $\beta_s = (\beta_{cs}, \beta_{ns})$ is a vector that describes how much each skill type determines outcome y_s.

[6] Students may possess many types of cognitive and non-cognitive skills. The key point is that the extention relaxes the assumption that students are eithter high- or low-skilled, and permits the more realistic scenario in which students may be highly skilled on certain dimantions but deficient in other diamensions of skill.

[7] The assumption that student ability and teacher quality are additively separable is common to all value-added models. Empirical tests have found little evidence against the additive model.

[1] $$y_{sij} = \alpha_{ij}'\beta_s = (\upsilon_i + \omega_j)'\beta_s,$$

There is a longer-run outcome y_l that policymakers care about (such as high-school graduation, college going, or earnings) but cannot be measured contemporaneously. The longer-run outcome is also a function of a student's stock of cognitive and noncognitive skill. The long-run outcome is $y_{lij} = \alpha_{ij}'\beta_l + \varepsilon_{lij}$, where ε_{lij} is random error and $\beta_{cl} \times \beta_{nl} \neq 0$.

Teachers' Effects**:** Teachers affect student outcomes only through their effects on students' accumulated skills. From [1], teacher j's effect on any outcome y_z, where z={*s*,*l*}, is a weighted average of her effect on each dimension of student ability, and is given by [2].

[2] $$\theta_{zj} = \omega_j'\beta_z.$$

Claim 1: *Teachers can systematically improve non-test score outcomes and long-run outcomes without improving test scores.*

To show that this can be true, consider this stylized example. There are two 9th grade outcomes: test scores (y_1) and another outcome (y_2). Suppose test scores are only a function of cognitive skill (i.e. $\beta_{1c} \neq 0$ and $\beta_{1n} = 0$) and the other outcome is only a function of noncognitive skill (i.e. $\beta_{2c} = 0$ and $\beta_{2n} \neq 0$). Consider teachers who have no effect on cognitive skill but do affect students' noncognitive skill (i.e. $\omega_{cj} = 0$ and $\omega_{nj} \neq 0$). These teacher's effect on test scores will be $\theta_{1j} = \omega_{cj}\beta_{1c} = 0$, these teacher's effect on the non-test score outcome will be $\theta_{2j} = \omega_{cj}\beta_{2c} \neq 0$, while their effects on the longer run outcome will be $\theta_{lj} = \omega_j\beta_l \neq 0$.

Claim 2: *One can better predict a teacher's effect on long-run outcomes using multiple short-run outcomes that reflect a different mix of both ability types than using test scores alone.*

Consider two 9th grade outcomes, test scores (y_1) and another outcome (y_2), and a long run outcome (y_l). The best linear unbiased estimate of the teacher effect on long-run outcome (y_l) based on the effect on test scores is $\gamma\theta_{1j}$, where $\gamma = cov(\theta_{lj}, \theta_{1j})/var(\theta_{1j})$. It is straightforward to show that the variation in a teacher's effect on the long run outcome (θ_{lj}) unexplained by her effect on test scores (θ_{1j}) is a linear function of her quality vector $\ddot{\theta}_{lj} = f(\omega_{cj})$.[8] Similarly, the variation

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

in a teacher's effect on the additional outcome (θ_{2j}) unexplained by her effect on test score (θ_{1j}) is also a linear function of the same quality vector $\ddot{\theta}_{2j} = g(\omega_{cj})$. Teacher effects on y_2 will increase the explained teacher-level variability in the long-run outcome *iff* $cov(f(\omega_{cj}), g(\omega_{cj})) \neq 0$.[9] Because both $f(\omega_{cj})$ and $g(\omega_{cj})$ are functions of the same vector ω_j, it follows that $cov(f(\omega_{cj}), g(\omega_{cj})) \neq 0$ so that teacher effects on y_2 will increase the explained teacher-level variability in the long-run outcome. I present evidence of this in Section VI. Intuitively, if an additional outcome reflects a different mix of skills from that measured by test scores, teacher effects on that additional outcome may explain variation in her effect on the long-run outcome that are not explained by her effect on test scores.[10] It is important to stress that *this result does not require that the additional outcome be unrelated to test scores, but the much weaker condition that there is meaningful variation in the other outcome that is unrelated to test scores*.

III Data and Relationships between Variables

I seek to estimate the effect of 9th grade teachers on test scores and behaviors, and explore whether these estimates predict teacher effects on longer-run outcomes. I use data on all public school students in 9th grade in North Carolina between 2005 to 2012 from the North Carolina Education Research Data Center. The data include demographics, transcript data, test scores in grades 7 through 9, and codes linking student test scores to the teacher who administered the test.[11] I focus on students who took the Algebra I or English I courses (the two courses for which standardized tests have been consistently administered over time). Over 90 percent of all 9th graders take at least one of these courses, so the sample is representative of 9th graders. To avoid bias that would result from teachers having an effect on students repeating 9th grade, I use only the first observation of 9th grade repeaters.[12] Summary statistics are presented in Table 1.

These data cover 537,241 ninth grade students in 676 secondary schools, 5,049 English I teachers, and 4,703 Algebra I teachers. The gender split is roughly even. The sample is 59.3 percent white, 25.9 percent black, 7.2 percent Hispanic, and 2 percent Asian. Regarding the highest

[9] See Appendix 4 for a formal proof of this statement.
[10] This could also be 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 used in this paper.
[11] Because the teacher identifier listed is not always the student's teacher, I use an algorithm to ensure high quality matching of students to teachers. I detail this in Appendix 1.
[12] Results that exclude 9th grade repeaters entirely are essentially unchanged.

education level of students' parents (i.e., the highest level of education obtained by either of the student's two parents), 6.7 percent were below high school, 39.6 percent had a high school degree, 15.1 percent had a junior college or trade school degree, 22.5 percent had a four-year college degree or greater, and 6.6 percent had an advanced degree (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 each testing year. Test scores in are higher than average because the sample of 9th graders successfully matched to their classroom teacher are slightly higher achieving on average.[13]

Informed by studies that have used behaviors as proxies for "soft" skills (e.g. Lleras 2008, Bertrand and Pan 2013, Kautz 2014), I proxy for noncognitive skill using non-test-score outcomes available in the data; the log of the number absences in 9th grade, whether the student was suspended during 9th grade, 9th grade grade point average (all courses), and whether they enrolled in 10th grade on time. These outcomes are strongly associated with well-known psychometric measures of noncognitive skills including the "big five" and grit.[14] Following Heckman, Stixrud, and Urzua (2006), I use a factor model to create a single index of these behavioral outcomes and to account for measurement error in each of them. 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 behavioral factor.[15] While test scores will certainly reflect *some* of the same skills as those measured by the factor, the variation in this factor that is unrelated to test scores may serve as a proxy for a set of skills that may go largely unmeasured by standardized tests.[16]

As one might expect, the behavioral factor and test scores are positively correlated. The behavioral factor has a correlation of 0.51 with Algebra scores and 0.50 with English scores. This

[13] Also, test scores in 7th and 8th grade are higher than the average because (a) the sample is based on those higher achievers who remained in school through 9th grade, and (b) I use the most recent 8th or 7th grade score prior to 9th grade which will tend to be higher for repeaters. Algebra I and English I scores are also slightly above zero because the classrooms that can be well matched to teachers have slightly higher performance than average.

[14] Low agreeableness and high neuroticism are associated with more absences, externalizing behaviors, juvenile delinquency, and lower educational attainment (Lounsbury, et. al. 2004; Barbaranelli, et. al. 2003; John, et. al. 1994; Carneiro et. al. 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).

[15] I estimated a factor model on the behavioral outcomes and then computed the unbiased prediction of the first underlying factor. This predicted factor was computed using the Bartlett method, however the results are robust to other methods. The predicted factor is Factor = -0.45*absences -0.35*suspended+0.64*GPA +0.57*on time in 10th grade. See Appendix 2 for the correlations between the 9th grade outcomes.

[16] 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, & Parrott, 2000; Brookhart, 1993). As such, grades reflect a combination of skills only some of which may be measured by test scores.

is consistent with the commonsense view that, in general, successful students tend to score well on tests and also be relatively well behaved. Analysis of Variance (ANOVA) reveals that about 75 percent of the variation in the behavioral factor is unrelated to test scores. If this 75 percent reflects real skills, then the factor may contain information that can be used to identify teachers that improve longer-run outcomes. The extent to which teachers have causal effects on these outcomes, and the extent to which teacher effects on these outcomes measure skills that are unmeasured by test scores but reflected in longer-run outcomes are the empirical questions tackled in Section VI.

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, effects observed on both outcomes cannot be due to changes in private school or out-of-state enrollment. Data are collected on high school GPA at graduation, SAT taking, and reported intentions to attend a four-year college upon graduation (2006 through 2011 cohorts). Roughly 4.2 percent of 9th graders subsequently dropped out of school, while 82.7 percent graduated from high school. The remaining 11 percent either transferred out of the North Carolina school system or remained in school beyond the expected graduation year. Roughly 47.3 percent of 9th graders took the SAT by 12th grade, and 27 percent intend to attend a four-year college.

III.1 *Motivating the use of behavioral outcomes as a proxy for skills*

To further motivate the use of behaviors as a proxy for skills that may not be well-measured by test scores, this section presents evidence that increases in test scores and behaviors are independently associated with better longer-run outcomes (Table 2). While the patterns presented here are descriptive, Section VI presents relationships that can be interpreted causally. I regress longer-run outcomes on GPA, absences, being suspended, on-time grade progression, and test scores (all measured in 9th grade). To remove the influence of socio-demographics, all models include controls for parental education, gender, ethnicity, English and math test scores in 7th grade and 8th grade, repeater status in 8th grade, absences in 8th grade, out of school suspension in 8th grade, and include indicator variables for each secondary school. Columns 1 and 2 show that higher test scores in 9th grade predict less dropout and more high-school graduation. However, they also show that the non-test score outcomes in 9th grade predict variability in these longer-run outcomes *conditional on test scores*. As expected, higher GPAs and on-time grade progression predict lower

dropout rates and more high-school graduation. Similarly, increased suspensions and absences predict higher dropout and lower high-school graduation. For both outcomes, one rejects the hypotheses that the non-test score outcomes in 9th grade have no predictive power for the longer-run outcomes conditional on test scores at the one percent level.

Using the behavioral factor that combines the non-test score outcomes into a single noncognitive factor to account for measurement error, columns 3 and 4 show that for both longer-run outcomes a standard deviation (σ) increase in the behavioral factor is associated with sizeable improvements conditional on test scores (results are similar using Math or English test scores). To summarize test scores with a single variable and to account for measurement error in test scores, I create a test-score factor that is a weighted average of algebra and English scores in 9th grade. While a 1σ increase in the test-score factor is associated with a 1.6 percentage point decrease in dropout, a 1σ increase in the behavioral factor is associated with a 4.59 percentage point decrease in dropout. Similarly, while a 1σ increase in the test-score factor is associated with a 2.95 percentage point increase in high-school graduation, a 1σ increase in the behavioral factor is associated with a 15.4 percentage point increase. Importantly, the patterns are similar for the predictors of college going, high-school grade point average at graduation, SAT-taking, and college plans. Across all the longer-run outcomes, increases in the behavioral factor are associated with large improvements conditional on test-scores. This suggests that the behavioral factor may be a good predictor of longer run outcomes above and beyond effects predicted by test scores.

To further validate the behavioral factor, in Appendix 3 I replicate the patterns in Table 3 using nationally representative survey data — the National Educational Longitudinal Survey of 1988 (NELS-88). I also demonstrate that, in the survey data, the behavioral outcomes predict educational completion, crime, and labor market outcomes conditional on test scores. Psychometric measures of noncognitive skills have been found to be particularly important at the lower end of the earnings distribution (Lindqvist & Vestman, 2011; Heckman, Stixrud, & Urzua, 2006). To see if this is also true for the behavioral factor, I estimate the marginal effect of the factor on log earnings at different points in the earnings distribution (Appendix 3). Similar to psychometric measures of noncognitive skills, the behavioral factor has much larger effects at the lower end of the earnings distribution conditional on test scores — further evidence that the behavioral factor captures noncognitive skills not well-measured by scores on standardized tests.

IV Empirical Strategy

This section outlines the "modified" value-added model used to estimate teacher effects on student test-scores and behaviors in 9th grade. The estimated effects on 9th grade outcomes will then be used as predictors of longer-run outcomes. To derive a statistical model from the model in Section II, I introduce randomness to student outcomes in 9th grade. Each outcome y_z for student i with teacher j is a function of student skill at the end of 9th plus a random error leading to [3].

[3] $$y_{zij} = (\upsilon_i + \omega_j)'\beta_z + \varepsilon_{zij} = \upsilon_i'\beta_z + \omega_j'\beta_z + \varepsilon_{zij}.$$

Cross multiplying out the first term and substituting equation [2] leads to [4].

[4] $$y_{zij} = \upsilon_{ci}\beta_{zc} + \upsilon_{ni}\beta_{zn} + \theta_{jz} + \varepsilon_{zij}.$$

Equation [4] shows that conditional on students' incoming endowments of cognitive and noncognitive ability υ_{ci} and υ_{ni}, one can identify the average effect of each teacher on any outcome, θ_{zj}. The identifying assumption in value-added model with one-dimensional ability is that lagged test scores are a proxy for incoming student ability (Todd and Wolpin 2003). With two dimensions of ability, including lagged values of *any* two linearly independent outcomes is sufficient to proxy for students' incoming skills in both dimensions.[17] All models include lagged values of *five* outcomes; math scores, English scores, repeater status, suspensions, and attendance. Lagged GPA is not included because those data are not available in middle school. However, five lagged outcomes are more than sufficient to proxy for two dimensions of skill. Moreover, to assuage any lingering concerns about using GPA as an outcome without conditioning on lagged GPA, Appendix 6 shows that the results are robust to excluding GPA from the analysis entirely.

Even though lagged outcomes are powerful controls for incoming student characteristics, to account for other sources of sorting and differences in schooling inputs, I employ three empirical approaches suggested by the teacher quality literature *simultaneously*; I control for lagged peer outcomes as suggested in Protic at. al. (2013), the number of honors courses taken as suggested in both Harris and Anderson (2012) and Aaronson et al (2007), and I also include fixed effects for the student's academic school track as suggested in Jackson (2014). The academic school track is the unique combination of the ten largest academic courses, the level of Algebra I taken, and the level of English I taken *in a particular school*.[18] As such, only students at the same school who

[17] See Appendix 4 for a formal proof.

[18] 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 5 for further discussion of tracks.

also take the same academic courses, level of English I, and level of Algebra I are in the same school track.[19] I refer to the academic school track as "track" for the remainder of the paper. The validity of teacher effects based on value-added models has been demonstrated using experimental variation in several contexts (Kane and Staiger 2008; Kane, McCffrey, Miller and Staiger 2013). However, to assuage lingering concerns of bias, I implement empirical tests suggested by Chetty, Rockoff and Freidman (2014a), and find no evidence of bias due to selection or tracking.

Including all the aforementioned conditioning variables, I follow convention in the value-added literature and model outcome *z* of student *i* with teacher *j* in year *t* with equation [5].

[5] $$y_{zijt} = \Omega_z X_{it} + e_{zit}$$

Here, X_{it} denotes all observable student and class characteristics to account for tracking, sorting, and incoming student ability; these include incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), the number of honors courses taken during 9th grade, and indicator variables for each track. If one removes the influence of the observable predictors, one is left with $e_{zijt} = y_{zijt} - \Omega_z X_{it}$. This residual error is comprised of the effect of the teacher θ_{zj}, a random classroom-level shock ε_{zc}, and an idiosyncratic student-level shock ε_{zi}, such that $e_{zijt} = \theta_{zj} + \varepsilon_{zc} + \varepsilon_{zi}$. The average of these student level residuals for a given teacher ($\bar{e}_{zj}$) is an unbiased estimate of the teacher's effect on outcome *z* under the identifying assumptions.

Even though $\bar{e}_{zj}$ is an unbiased estimate of a teacher's effect, to avoid endogeneity, one should not estimate teacher effects using the same students among which longer-run outcomes are being compared. Accordingly, I follow Chetty et. al. (2014a) and *predict* how much each teacher improves student outcomes in a given year based on her performance in *other* years (based on 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 effects that persists over time. A leave-year-out estimate for teacher *j* in year *t* is the teacher's average residuals based on all other years of data as in [6].

[19] 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 same courses but taking Algebra or English at different levels are in different school tracks. Because many students pursue the same course of study, less than one percent of all students are in singleton tracks, 82 percent of students are in tracks with more than 20 students, and the average student is in a school track with 175 other students.

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

Because $\hat{\theta}_{zj,-t}$ is estimated with noise, researchers use the raw means to form empirical Bayes (or Shrinkage) estimates of teacher quality (Staiger and Kain 2008; Chetty at al 2014a; Gordon, Kane, and Staiger, 2006). Because, this is also the approach used by districts for policy purposes, I employ this approach. This approach models the estimation error in each teacher's raw mean and adjusts (or shrinks) noisier estimates towards the grand mean (in this case zero). The resulting leave-year-out Empirical Bayes estimate used for teacher *j* is described by [7].[20]

[7] $$\hat{\mu}_{zjt} = \hat{\theta}_{zj,-t}\left[\frac{\sigma^2_{\theta zj}}{\sigma^2_{\theta zj}+(\sigma^2_{\varepsilon zc}+\sigma^2_{\varepsilon_{zi}}/n)/t-1}\right] \equiv \hat{\theta}_{zj,-t}\lambda_{zj}.$$

This empirical Bayes estimate for each teacher's effect is the leave-year-out teacher-level mean ($\hat{\theta}_{zj,-t}$) multiplied a by λ_{zj}, an estimate of its reliability. As a result, less reliable 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. Because Empirical Bayes estimates explicitly account for noisiness in the estimates, they tend to be better predictors of outcomes when used as covariates in a regression setting. See Staiger and Kain (2008), Morris, (1983) and Reardon and Raudenbush (2009) for discussion of this approach. To examine whether teacher effects on test scores and the behavioral factor predict effects on longer-run outcomes, I use the estimates from [7] as predictors of the longer-run outcomes.

V Effects on Test Scores and Non-Test Score Outcomes in 9th Grade

Before presenting teacher effects on longer-run outcomes, I examine the magnitudes of the teacher effects on 9th grade outcomes. I follow Kane and Staiger (2008) and for each outcome, use the covariance between mean classroom-level residuals for the same teacher as a measure of the variance of the persistent component of teacher effects ($\hat{\sigma}^2_{\theta_{zj}}$).[21] The estimated variances for all 9th

[20] This is the same as equation (9) from Chetty et. al. (2014a) and equation (5) in Kane and Staiger (2008). Following the literature, the parameters $\sigma^2_{\theta zj}$, $\sigma^2_{\varepsilon zc}$, and $\sigma^2_{z\varepsilon_i}$ are estimated using the covariance of the error terms across classrooms under the assumption that $cov(\theta_{zj}, \varepsilon_{zc}) = cov(\theta_{zj}, \varepsilon_{zi}) = cov(\varepsilon_{zi}, \varepsilon_{zc}) = 0$. Under this assumption, var(e_{zijt}) $= \sigma^2_{\varepsilon zi} + \sigma^2_{\varepsilon zc} + \sigma^2_{\theta zj}$, and $cov(\bar{e}_{zjct,}, \bar{e}_{zjc\prime,-t}) = \sigma^2_{\theta zj}$ where $\bar{e}_{zjct}$ is the average residual for classroom *c* for teacher *j* in year *t* and $\bar{e}_{zjc\prime,-t}$ is the average residual for classroom *c'* for teacher *j* not in year *t*. $\sigma^2_{\varepsilon_{zi}}$is estimated using the variance of the student level residuals within classrooms, and $\sigma^2_{\theta zj}$ is estimated using the covariance of classroom-level mean residuals for the same teacher in *different* years. Finally, $\sigma^2_{\varepsilon zc}$ is estimated as the variance of the total residual, var(e_{zijt}), minus the estimates of $\sigma^2_{\varepsilon_{zi}}$ and $\sigma^2_{\theta zj}$.

[21] I compute mean residuals ($\bar{e}_{zct}$) for each classroom. Then I link every classroom-level mean residual and pair it

grade outcomes are presented for each subject in Table 3.

The standard deviation of the Algebra teacher effects on Algebra test scores is 0.0654σ. This indicates that having an Algebra teacher at the 85th versus 15th percentile of effects on algebra test scores would increase algebra scores by roughly 0.13σ. To put this into perspective, the partial correlations in Table 2 imply that this would be associated with being 0.38 percentage points more likely to graduate from high school. Looking to the non-test score outcomes, having an Algebra teacher with estimated effects at the 85th versus15th percentile reduces the likelihood of being suspended by 2.48 percentage points, reduces absences by 4 percent, increases GPA by 0.034 grade points, and increases on-time grade progression by about 2 percentage points. Combining the non-test-score outcomes into a single variable, the standard deviation of Algebra teacher effects on the behavioral factor is 0.04σ, so that having an Algebra teacher at the 85th versus 15th percentile of effects on the factor would increase the behavioral factor by 0.08σ. The partial correlations in Table 2 suggest that this would be lead to a 1.2 percentage point increases in the likelihood of high-school graduation. Given the large benefits to graduating from high school, if effects on the longer-run outcomes are similar to those implied by the partial correlations, the magnitudes of the teacher effects on both the test-score and non-test score outcomes are economically meaningful.

Patterns for English teacher are largely similar to those for Algebra teachers. However, as has been found in other settings, teacher effects on English scores are smaller than those on math scores. The standard deviation of English teacher effects on scores is 0.03σ so that having an English teacher at the 85th percentile of effects on English test scores versus the 15th percentile would raise English scores by 0.06σ. Summarizing the non-test-score effects, the standard deviation of English teacher effects on the behavioral factor is 0.03389σ -- an effect size on behaviors that is similar to those for Algebra teachers. The patterns presented in Table 3 indicate that there may be economically meaningful variation in outcomes across teachers that persists across classrooms. Whether this variation can be well-measured for individual teachers, and whether estimated effects on different outcomes measure different skills are explored below.

V.2 *Relationship between Teacher Effects across 9th Grade Outcomes*

To gain a sense of whether teachers who improve test scores also improve other outcomes,

with another random classroom-level mean residual for the same teacher and compute the covariance of these mean residuals. As discussed in footnote 20 the covariance of mean residuals within teachers but across classrooms is a consistent measure of the true variance of persistent teacher quality. I replicate this calculation 1000 times and take the median of the estimated covariance as the parameter estimate.

Table 4 presents the raw correlations between the estimated teacher effects on the different outcomes in 9th grade where the data for both Algebra and English teachers are combined. Teachers with higher test score effects are associated with better non-test score outcomes, but the relationships are weak. The correlations between test score effects and effects on being suspended or absences are both below 0.1. The test score effects are somewhat more highly correlated with GPA (r=0.1933) and on-time grade progression (r=0.1315), but not strongly so. The correlation between teacher effects on test scores and teacher effects on the behavioral factor is a modest 0.164. This indicates that less than 3 percent of variability in teacher effects on the behavioral factor is associated with teacher effects on test scores, and *vice versa*. This indicates that many teachers who improve test scores may have small effects on non-test-score outcomes and *vice versa*. This may suggest that test score effects measure effects on certain skills, and teacher effects on the behavioral factor measure effects on a largely *different* but potentially important set of skills.

To explore further whether teacher effects on test scores and the behavioral factor may measure different sets of skills, I regress test scores and the behavioral factor on the estimated teacher effects for those two outcomes. If effects on the behavioral factor and test scores measure distinct dimensions of skills, then predicted teacher effects on test scores should predict test scores but *not* the behavioral factor, and predicted teacher effects on the behavioral factor should predict the behavioral factor but not test scores. However, if they measure the same set of skills, then predicted teacher effects on both outcomes should predict changes in both outcomes.

To implement this test I estimate the following regression model where all variables are defined as in [8] and $\hat{\mu}_{test,jt}$ and $\hat{\mu}_{behaviour,jt}$ are the leave-year-out Empirical Bayes teacher effect estimates on test scores and the behavioral factor, respectively.

[8] $$y_{zijt} = \Omega_z X_{it} + \delta_{z1} \cdot (\varphi_1 \hat{\mu}_{test,jt}) + \delta_{z2} \cdot (\varphi_2 \hat{\mu}_{behavoir,jt}) + \nu_{zit}.$$

For ease of interpretation, the estimated teacher effects are multiplied by scaling factors φ_1 and φ_2 so that the coefficients δ_1 and δ_2 identify the effect of increasing the teacher effect on test scores and the behavioral factor, respectively, by one standard deviation (i.e. going roughly from a teacher at the median of the effect distribution to one at the 85 percentile).[22] Data for both subjects

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

[a] $$y_{zijt} = \beta_z X_{it} + \pi_z \cdot \hat{\mu}_{zjt} + \nu_{zit}$$

The scaling factor is $\varphi_1 = \hat{\pi}_z / \hat{\sigma}_{\theta_{zj}}$, where $\hat{\pi}_z$ is the coefficient estimate from [a] and $\hat{\sigma}_{\theta_{zj}}$ is the estimated standard deviation of the true teacher effects on outcome z described in Table 3. It is straightforward to show that the coefficient on the rescaled teacher effect for outcome z on outcome z will be $\hat{\sigma}_{\theta_{zj}}$. The coefficient on the rescaled teacher effect

are stacked and the results are presented for both subjects combined. Section VI presents results separately by subject. Standard errors are adjusted for clustering at the teacher level.

Table 5 presents the regression coefficients on the rescaled leave-year-out empirical Bayes teacher effect estimates. As one might expect, out-of-sample estimated teacher effects on a particular outcome have large statistically significant effects on that outcome. Column 1 shows that increasing teacher test score value-added (across both subjects) by one standard deviation increases test scores by 0.05σ (p-value<0.01). As one might expect, this is between the estimated standard deviations for Algebra teachers (0.065) and that for English teachers (0.030). Looking at behaviors, Column 5 shows that increasing the teacher effect on behaviors by one standard deviation increases the behavioral factor by 0.0338σ (p-value<0.01). Consistent with teacher effects on these two outcomes measuring different dimensions of skill, Column 2 shows that increasing teacher effects on behaviors has no effect on test scores, and Column 5 shows that increasing teacher effects on test scores has no effect on behaviors. The coefficients are both small and neither is statistically significant. Given that the behavioral factor and test scores had a moderate to weak positive correlation, one would expect that effects on one outcome would predict improvements in the other outcome. However, the point estimates suggest that teacher effects on these two outcomes reflect distinct dimensions of skills that are largely orthogonal

As indicated in the model, variability in outcomes associated with individual teachers that is unexplained by effects on test scores may reflect other unmeasured skills. If this is so, and the behavioral factor is a reasonable proxy for these other skills, then teacher effects on the behavioral factor might explain variability in teachers' ability to improve long-run outcomes that is not measured by effects on test scores. Section VI investigates this directly.

VI Predicting Longer Run Effects with Short Run Effects

The main longer-run outcomes under study are measures of high school completion. While the relationships in Table 2 *suggest* that teachers who improve behaviors may improve longer-run outcomes, this section directly tests whether teachers who increase the behavioral factor *cause* improved longer-run outcomes (conditional on test score effects). I estimate equation [7] where the outcomes are measures of high school completion; whether the student subsequently dropped

on outcome z has the convenient interpretation of being the marginal effect of increasing the teacher effect on outcome z by one standard deviation (i.e. going roughly from a median teacher to one at the 85 percentile).

out of secondary school by 12th grade, and whether they graduated from high school by 12th grade.

As before, the coefficients on rescaled teacher effects on test scores and the behavioral factor represent the effect of increasing the teacher effect on test scores and the behavioral favor by one standard deviation, respectively. To quantify the increase in the ability to predict variability in teacher effects on the longer-run outcome by adding effects on the behavioral factor, I estimate [7] both with and without the effects on behaviors, and I compute the percentage increase in the predicted variability of the teacher effects on the long-run outcome.[23] The results are presented for both subjects in Table 6. Standard errors are adjusted for clustering at the teacher level.

Column 1 presents the effect of increasing test score value added on high-school graduation when the effect on behaviors is not included. On average, one standard deviation higher test score value added leads to a 0.138 percentage point increase in high-school graduation (p-value<0.05). To put this into perspective, the partial correlations between test scores and graduation in Table 2 show that increasing test scores by 1 standard deviation would increase the likelihood of graduating high school by 2.9 percentage points. Given that one standard deviation of the teacher effect is roughly 0.05σ (in student units), the partial correlations imply that a one standard deviation increase in teacher test score value-added would increase high-school graduation by 2.9*0.05=0.145 percentage points. This is very close to the estimated magnitudes—suggesting that the results are reasonable. It is also helpful to compare the estimates to those from Chetty et. al. (2014b). Their estimates indicate that a teacher who raises test scores by 0.05 standard deviations would increases college going by 0.28 percentage points.[24] The estimates in Table 6 are smaller than those implied by Chetty et al (2014b), but they are of a similar order of magnitude.

Column 2 presents the impact of teacher effects on the behavioral factor on high-school graduation when test-score effects are not included. On average, students in class with a teacher with one standard deviation higher behavioral factor effect are 0.78 percentage points more likely to graduate high school (p-value<0.01). This is similar to what one would expect based on the partial correlations between the behavioral factor and high-school graduation, which is 0.52 percentage points. Column 3 presents the effect on high-school graduation of teacher effects on

[23] I compute the variance of the fitted values for each teacher from [7]. In models without the effect on the behavioral factor this is $a = var(\hat{\delta}_1 \cdot (\varphi_1 \hat{\mu}_{test,jt}))$, and in models with teacher effects on both, this is $b = var[\hat{\delta}_1 \cdot (\varphi_1 \hat{\mu}_{test,jt}) + \delta_2 \cdot (\varphi_2 \hat{\mu}_{behavoir,jt})]$. The percentage increase in the explained variability from also including the teacher effect on the behavioral factor (versus using test score value added alone) is 100(a/b-1).

[24] They find that a teacher who raises test score value added by one standard deviation increases test scores by 0.15σ and college going by 0.82 percentage points.

both the behavioral factor and test scores. Given that the two effects are largely uncorrelated, the point estimates remain largely unchanged. Increasing test score value added by 1 standard deviation increases high-school graduation by roughly 0.1 percentage points, and increasing a teacher's behavioral effect by 1 standard deviation increases high-school graduation by roughly 0.72 percentage points. Comparing the teacher-level variability on high-school graduation from the fitted models with both effects to those using only test score value added, including teacher effects on the behavioral factor increases the explained variability of teacher effects on graduation by 249% percent – i.e. more than triples the identifiable teacher effect on high-school graduation.

While these effects may seem modest, consider the following back-of-the-envelope calculation. Increasing a teacher's behavioral effect by 1 standard deviation increases high-school graduation by 0.72 percentage points, on average. The average teacher has 54.5 students a year. According to the Bureau of Labor Statistics (2016), completing high school is associated with $220 higher weekly earnings – an annual difference of $11,000. Assuming this difference is causal, increasing the likelihood of graduating by 0.72 percentage points would increase annual earnings by roughly $80 per year per student. This figure multiplied by 54 students is $4276 higher cohort earnings each year. If we assume that this increase stays the same each year for 40 years, (at a 7% discount rate) this translates into $62,278 in present discounted lifetime earnings per year of students taught. If one made the conservative assumption that half of the $220 increase in weekly earnings is due to selection, the effect sizes translate into $30,601 in lifetime earnings per year of students taught. In sum, under most reasonable assumptions regarding the economic benefits of completing high school, the estimated effects are economically important.

The other measure of high school completion is high-school dropout. High-school dropout is notoriously difficult to measure (Tyler and Lofstrom 2009) so that the estimated effects will likely be muted. However, it is helpful to show that the same basic patterns that hold for high-school graduation also hold for high-school dropout. Column 4 shows that, on average, a one standard deviation increase in teacher test score value added reduces the likelihood of dropout by 0.06 percentage points (p-value<0.1). This point estimate is smaller than the effect on graduation, but it is also noisier so that one cannot reject that the two estimates are the same in a statistical sense. Column 5 shows that, on average, a one standard deviation increase in the teacher effect on behaviors reduces the likelihood of dropout by 0.517 percentage points (p-value<0.01). This is similar to the point estimates for high-school graduation – one cannot reject that the effect on the

two outcomes is the same at traditional levels of statistical significance.

Column 6 presents the effect on high-school dropout of increasing the teacher effects on both the behavioral factor and test scores. Increasing test score value added by 1 standard deviation decreases dropout by 0.047 percentage points, and increasing a teacher's behavioral effect by 1 standard deviation decreases dropout by 0.498 percentage points on average. Including teacher effects on the behavioral factor increases the explained variability of teacher effects on dropout by 527% percent. This very large *relative* increase reflects the fact that test score value added does not appear to be a very strong predictor of dropout in the full model. The similarity of the pattern of results across the two measures of high school completion supports the idea that the estimated effects are real and reflect real changes in human capital acquisition. While the increases in the explained variation may seem large, they are consistent with Chamberlain (2013) who finds that test score effects account for less than one fifth of the overall effect of teachers on college-going. However, if teachers have effects on 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.

To explore the possibility that the estimated effects are driven by any single outcome Appendix 6 present results where I use teacher effects on each behavioral outcome individually. For both outcomes, the teacher effects on the individual behavioral outcomes have the expected sign and some of them are statistically significant. Because lagged GPA is not included a conditioning variable, to ensure that the GPA variable is not the sole driver of the pattern of results, I demonstrated that the results are robust to using teacher effects on a factor that excludes the GPA variable entirely. In sum, Appendix 6 show that the effects on no single outcome is driving the effect on longer-run outcomes, and that it is the shared variability across the behavioral outcomes (which I posit is due to noncognitive skills) that drives the key results.

VI.1 Addressing Selection

Rothstein (2009) raises the concern 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. In response, Kane and Stagier (2008), Kane, et al (2013), Chetty et. al. (2014b), and Backer-Hicks et al (2015) all show that, in several contexts, teacher value added estimates exhibit no appreciable bias in experimental and quasi-experimental data. However, it is important to present evidence that selection does not drive the results in the current context.

To this aim, I first implement a test for selection on observables (Appendix 7). I show that

conditional on 8th grade outcomes and controls for tracks, teacher effect estimates are unrelated to predicted dropout and predicted graduation (weighted indices of parental education, 7th grade math scores, 7th grade reading score, gender, and ethnicity). To test for selection on unobservables within school track 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 the estimated teacher effects merely capture selection within school cohorts, then the arrival of a teacher who increases the average predicted teacher effect for a cohort but has no effect on real teacher quality or student outcomes should have no effect on average student outcomes for that cohort. Conversely, if the predicted effects are real, differences in average predicted teacher quality *across* cohorts (driven by changes in teaching personnel within schools over time) should be associated with similar differences across cohorts in average cohort-level outcomes as the same difference in estimated teacher quality across individual students *within* cohorts. I test for these patterns empirically (Appendix 7), and find that for both longer-run outcomes, results using the clean variation *across* cohorts and those based on the potentially endogenous variation *within* cohorts are similar. Also, Appendix 8 Table 1 shows that the results are robust to including school-by-year fixed effects and Appendix 8 Table 2 shows that the results are robust to including behaviors in 7th grade. Consistent with other studies, I find little evidence of selection on observables with a sufficiently rich set of controls, and I can rule out selection on unobservables as the driver of the observed patterns.

VI.2 Effects by Subject

The results thus far have analyzed English and Algebra teachers together. I relax this restriction and show effect for English and algebra teachers separately. This is accomplished by interacting the estimated teacher effects with indicators for the subject and including these interactions in the regression model. The estimates effects are presented in Table 7. Column 1 shows the estimated effect on test scores and the behavioral factor. As expected, test score effects predict test scores and the effects are larger for Algebra teachers (0.072σ) than for English teachers (0.033σ). While both test score effects have statistically significant effects on test scores at the 1 percent level, estimated effects on the behavioral factor have no effect on test scores in either subject. The results in Column 2 reveal some interesting differences across the subjects. Specifically, the predicted effects on the behavioral factor strongly predict the behavioral factor

for English teachers (p-value<0.01), but have no statistically significant effect for Algebra teachers (p-value>0.1). This suggests that while the estimated effects on behaviors capture strong persistent teacher-level effects in behaviors for English teachers, this is not so for Algebra teachers.

While explaining the difference across subjects is beyond the scope of this paper, research on classroom practices provides some guidance. Survey data reveal that high school math teachers typically follow a pre-specified math textbook while English teachers tend to tailor their courses by choosing texts and topics (Siskin 1991). Because English classes involve more classroom discussion than math classes (Siskin 1991), I conjecture that English teachers influence student motivation and aspirations by selecting texts that embody themes such as perseverance, hard work, and resilience, and then orienting discussions around these themes. Even though this is speculative, Lee (2007) studied an intervention that focused English instruction on identity and resilience themes embodied in literature readings. She found that the intervention was associated with positive changes on both psycho-social measures and also outcomes such as grades, and discipline – patterns that are consistent with my conjecture. Despite the reasons, the fact that English teachers affect a proxy for noncognitive skills helps explain a puzzle from Chetty et. al. (2014). They find that English teachers have smaller effects on test scores but larger effects on adult outcomes. If English teachers in their data also have effects on noncognitive skills, it could explain their result.

Having established that teachers in both subjects have real effects on test score but that only English teachers have appreciable effects on the behavioral factor, I now turn to the longer-run outcomes. As expected, Columns 3 and 4 show that English teachers effect on the behavioral factor predict large effects on dropout and graduation, while that for Algebra has no effect. Another interesting pattern is that the positive test-score effects on dropout and graduation appear to have been driven mostly by Algebra teacher effects on test scores. Taken together, the results imply that test score effects predict Algebra teacher effects on longer-run outcome while behavioral effects predict English teacher effects on the longer-run outcomes.

VI.3 Other Outcomes

While high-school dropout and graduation are the main long-run outcomes in this study, I also present effects of 9th grade teachers on whether a student took the SAT, whether they expressed intentions to attend a four-year college in a high-school graduation survey, and their high school GPA at graduation (Table 7). I focus attention on the teacher effects on the behavioral factor conditional on test score effects. English teacher effects on the behavioral factor predict

teacher effects on SAT taking, intentions to attend a four-year college, and high school GPA, conditional on test score effects. Specifically, conditional on teacher effects on test scores, an increase in the teacher effect on the behavioral factor by one standard deviation increases SAT taking by 0.485 percentage points (p-value<0.1), increases the likelihood of planning to attend a four-year college after high-school graduation by 1.2 percentage points (p-value<0.01) and increases high school GPA by 0.0159 points (p-value<0.01). Consistent with Algebra teachers having no systematic effect on the behaviors, increases in Algebra teachers' effects on the behavioral factor do not predict any changes in any of the longer run outcomes.

Chamberlain (2013) finds that test score effects account under one fifth of the overall effect of teachers on college-going – implying that one could increase the explained variability of teachers four-fold over that explained by test score value added. In fact, including the teacher effect on the behavioral factor increases the variance of the explained teacher effects by 231 percent for graduation, 603 percent for dropout, 71 percent for SAT-taking, 1063 percent for 4-year college intentions and 209 percent for high school GPA. In sum, teacher effects on the behavioral factor improve the ability to identify teachers who improve longer-run outcomes considerably, and the magnitude of the increased explained variability is in line with Chamberlain (2013).

VI.4 Testing for General Improvement in Skill

Readers may worry that the effects are mechanical and driven by grade inflation or differential reporting of bad behaviors. Because passing English I and Algebra I is required to graduate from high school, grade inflation could mechanically improve graduation and reduce dropout without any real skill improvement. There are three key patterns that show that the effects are not mechanical and not driven by teacher differences in grading or reporting bad behaviors.

(1) Table 7 documents effects of the behavioral factor on longer-run outcomes for English teachers but not Algebra. If the grade inflation hypothesis is true, then one should observe that English teachers have large effects on English I course grades while Algebra teachers have no effect on Algebra I course grades. The opposite is true. I test this by regressing the own-course grade on teacher effects on behaviors and all covariates from [7] for Algebra and English teachers. The coefficient on the behaviors effect for Algebra teachers on the Algebra course grade is 0.621 (p-value<0.01), and that for English teachers on the English course grade is 0.111 (p-value<0.01).

(2) There is no mechanical relationship between grades or reporting bad behaviors and taking the SAT, or four-year college plans. However, Table 7 shows that teacher effects on behaviors

predict effects on these 12th grade outcomes. This suggest that these effects are not mechanical.

(3) Teacher effects on outcomes such as suspensions (which are used to form the factor but have no mechanical association with graduation or dropout) independently predict effects on longer-run outcomes (Appendix 6). Moreover, teacher effects on the behavioral factor more strongly predict longer-run outcomes than effects on any individual behavior—showing that it is teacher effects on the *common* element of these behaviors that matter. If the effects were mechanical, teacher effects on GPA or on-time grade progression would be a *stronger* predictors of effects on longer-run outcomes than teacher effects on the factor. The opposite is true.

Having presented patterns that are inconsistent with mechanical effects or reporting effects, I now present patterns that are consistent with improvement in skills. If the positive effects of the behavioral factor are due to skill acquisition, teachers who raise noncognitive skills would affect their students' outcomes not only in their own classes (as discussed above) but also in *other* classes. I test for this by estimating equation [7] where the main outcomes are GPA during 9th grade, GPA in all *other* classes in 9th grade, and the overall high-school GPA. The coefficient on 9th grade GPA is 0.031 (p-value<0.01), that for GPA in all other 9th grade classes is 0.019 (p-value<0.01), and that for overall high-school GPA is 0.016 (p-value<0.01). Because the effects on *other* 9th grade courses and overall high-school GPA are both positive and almost identical, it demonstrates that a large part of the effect on outcomes is driven by improvements in *other* classes (contemporaneously and in subsequent years)— evidence of a general improvement in skills.[25] While none of the tests is dispositive on its own, the several pieces of evidence taken as a whole make a compelling case that the relationships are due to improvements in student skill.

VI.5 Possible Policy Uses of Effects on Behaviors

I briefly discuss potential policy uses for the behavioral outcomes. One policy use would be to identify those observable teacher characteristics associated with effects on the behavioral factor and select teachers with these characteristics. To determine the scope of this type of policy, I regress the behavioral factor on observable teacher characteristics while controlling for school tracks, year effects, and student covariates (Appendix 8 Table 3). While observable teacher

[25] In principle, this could also be due to teacher inflating grades, thus allowing student to focus on other activities and classes. However, the fact that the teacher's behavioral effects predict much larger changes in on own course grades Algebra than English, while the long run effects are only observed for English teachers is inconsistent with this hypothesis. I tis also possible that the behavioral factor measures improvement in reading skills that translate into improvements in all subjects. The fact that (a) one sees similar results using a factor that excludes course grades, and (b) teachers who raise the factor do not increase English test scores, makes this explanation unlikely.

characteristics have modest effects on test scores, none of the observable teacher characteristics — year of teaching experience, being fully certified, scoring well on teaching exams, having a regular license, and selectivity of a teacher's college (as measured by the 75th percentile of the SAT scores at the teacher's college) — have a strong statistically significant relationship with the behavioral factor. Looking only at English teachers, possessing an advanced degree is associated with high behavioral outcomes, and more years of experience and graduating from a selective college are *negatively* associated with the behavioral factor.[26] To summarize the effect of observable teacher characteristics, I took the fitted values from a regression predicting students' behavioral factor as a function of teacher characteristics. The fitted values are the predicted teacher effects on the factor based on observable characteristics. I then regressed dropout and graduation on these fitted values. The coefficient on graduation is positive and that on dropout is negative, indicating that teachers who have observable characteristics associated with improving the behavioral factor also tend to reduce dropout and increase high-school graduation. However, the coefficient estimates are small and not statistically significant. All in all, the observable teacher characteristics used in this research are not particularly good predictors of teacher effects on skills measured by the factor. Accordingly, using these particular observable teacher characteristics to identify excellent teachers may provide limited benefits. This does not preclude the use of more detailed teacher information to better predict teacher effects on a range of skills.

Another policy application is to incentivize teachers to improve the behavioral factor. However, because some of the outcomes that form the behavioral factor (such as grades and suspensions) can be "improved" by changes in teacher behavior that do not improve student skills (such as inflating grades and misreporting behaviors) attaching external stakes to the behavioral factor 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

[26] Teachers are often held accountable for student test scores but not behaviors. This creates incentives to improve test scores but not behaviors. As such, one might expect an experience gradient for test scores but not for the behavioral factor. 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.

provide teachers with incentives to improve the behaviors of students in their classrooms the *following* year (when teacher's influence may still be present, but they could no longer manipulate student behaviors). The idea of using follow-on courses to measure persistent teacher quality has been used in studies of college professors (e.g. Carrell and West 2010; Figlio et al 2015) and could be applied to younger grades. A final solution is to identify those teaching practices that lead to improvements in the behavioral factor and incentivize teachers to use these practices. Such approaches have been used successfully to increase test scores (Taylor and Tyler, 2012; Allen et al. 2011). In sum, the behavioral outcomes used in this study can be useful for policy.

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 skill. In the model, teachers may have effects on skills that affect long-run outcomes, are not reflected in test scores, but are reflected in *other* outcomes. I use an index of behaviors in 9th grade to proxy for noncognitive skills and find that 9th grade teachers have meaningful effects on both test scores and the behavioral factor. These test scores and behaviors appear to measure distinct skills, and teacher effects on behaviors explain significant variability in their effects on high-school graduation and dropout that are not captured by their test-score effects. Adding teacher effects on the behaviors more than doubles the predicted variability on longer-run outcomes for English teachers, but provides little additional explanatory power for Algebra teachers. The results highlight the fact using non-test score measures can be fruitful in evaluating teacher specifically and human capital interventions more broadly.

The results provide hard evidence of an idea that many believe to be true but has never been shown concretely – that teacher effects on test scores capture only a fraction of their effect on human capital. Despite the clear policy implications of this work, it is important to note that several of the non-test score outcomes employed in this paper are gameable. Despite this, there are a few feasible ways to use the behavioral factor for policy. However, further work may be needed to derive measures of noncognitive skills that are both informative and also difficult to manipulate by teachers. The patterns presented in this paper suggests that the gains in student skill and overall well-being from doing so may be considerable.

Tables and Figures

Table 1: *Summary Statistics of Student data*

Variable	Obs.	Mean	Std. Dev.	Std. Dev. within Schools	Std. Dev. within Tracks
Math z-score 8th grade	537241	0.225	(0.934)	(0.878)	(0.596)
Reading z-score 8th grade	537241	0.213	(0.938)	(0.894)	(0.669)
Repeat 8th grade	534411	0.006	(0.078)	(0.078)	(0.073)
Suspended (8th Grade)	537241	0.039	(0.193)	(0.191)	(0.181)
Absences (8th Grade)	537241	4.583	(5.615)	(5.553)	(5.216)
Student: Female	537241	0.504	(0.50)	(0.499)	(0.480)
Student: Black	537241	0.259	(0.438)	(0.392)	(0.359)
Student: Hispanic	537241	0.072	(0.258)	(0.253)	(0.241)
Student: White	537241	0.593	(0.491)	(0.436)	(0.399)
Student: Asian	537241	0.020	(0.141)	(0.138)	(0.132)
Parental education: Some High School	537241	0.067	(0.25)	(0.246)	(0.236)
Parental education: High School Grad	537241	0.396	(0.489)	(0.474)	(0.450)
Parental education: Trade School Grad	537241	0.016	(0.126)	(0.126)	(0.123)
Parental education: Community College Grad	537241	0.135	(0.341)	(0.339)	(0.329)
Parental education: Four-year College Grad	537241	0.225	(0.417)	(0.408)	(0.385)
Parental education: Graduate School Grad	537241	0.066	(0.249)	(0.242)	(0.228)
Parental education: Missing	537241	0.095	(0.293)	(0.279)	(0.265)
Number of Honors classes	537241	1.079	(1.380)	(1.234)	(0.572)
Algebra I z-Score (9th grade)	341334	0.029	(0.994)	(0.926)	(0.785)
English I z-Score (9th grade)	534695	0.044	(0.979)	(0.932)	(0.683)
Absences (9th Grade)	537241	3.430	(4.897)	(4.809)	(4.423)
Suspended (9th Grade)	537241	0.050	(0.219)	(0.216)	(0.204)
GPA (9th Grade)	537017	2.905	(0.827)	(0.772)	(0.569)
In 10^{th} grade on time	537241	0.901	(0.299)	(0.295)	(0.267)
Dropout (2005-2011 cohorts)	497315	0.042	(0.201)	(0.200)	(0.188)
Graduate (2005-2011 cohorts)	497315	0.827	(0.378)	(0.373)	(0.345)
Take SAT (2006-2011 cohorts)	441238	0.473	(0.499)	(0.483)	(0.408)
Intend to attend 4yr college (2006-2011 cohorts)	441238	0.270	(0.444)	(0.434)	(0.380)

Notes: These summary statistics are based on students who took the English I or the Algebra I exam and were linked to their classroom teacher. Incoming math scores and reading scores are standardized to be mean zero, unit variance for all takers in that year. The higher test score scores for 9th graders in the sample reflect the fact that those classrooms that could be matched to their teacher had slightly higher scores on average.

Table 2: *Predicting Longer Run Effect Using 9th Grade Outcomes*

	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 4yr
Grade Point Average (9th grade)	-0.0361**	0.0974**					
	[0.000917]	[0.00151]					
Log of # Absences (9th grade)	0.00721**	-0.0223**					
	[0.000398]	[0.000687]					
Suspended (9th grade)	0.0167**	-0.0470**					
	[0.00246]	[0.00371]					
On time in 10th grade	-0.0781**	0.321**					
	[0.00215]	[0.00337]					
Algebra z-score (9th grade)	-0.00753**	0.0152**					
	[0.000675]	[0.00114]					
Math z-score (9th grade)	-0.00465**	0.00584**					
	[0.000730]	[0.00121]					
Average Test Scores z-score			-0.0162**	0.0295**	0.216**	0.0830**	0.0592**
			[0.000922]	[0.00154]	[0.00200]	[0.00182]	[0.00154]
Behavioral factor z-score			-0.0459**	0.154**	0.382**	0.145**	0.0818**
			[0.000747]	[0.00106]	[0.00161]	[0.00100]	[0.000840]
Observations	305,185	305,185	305,185	305,185	238,279	273,088	273,088

Robust standard errors in brackets. ** p<0.01, * p<0.05, + p<0.1

In addition to including 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 7th and 8th grade, suspension in 8th grade, days absent in 8th grade and whether the student had repeated 8th grade.

Table 3: *Covariance Based Estimates of The Variability of Persistent Teacher Effects*

	Algebra Teachers						
	English Score	Algebra Score	Suspended	Log absences	GPA	In 10th on time	Behavioral Factor
Algebra Teachers Implied SD	0.01727	0.06542	0.01241	0.02054	0.01787	0.00950	0.04088
English Teachers Implied SD	0.03015	0.02670	0.00669	0.00000	0.02798	0.00893	0.03389

Notes: The estimated standard deviations are the estimated covariances in mean residuals from equation [5] 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 1000 times and report the median estimated covariance as my sample covariance. To construct the standard deviation of this estimated covariance, I pair each classroom with a randomly chosen classroom under a different teacher and estimate the covariance.

Table 4: *Correlations Between Estimated Teacher Effects*

	Teacher Effect: Test Score	Teacher Effect: Suspended	Teacher Effect: Absences	Teacher Effect: GPA	Teacher Effect: In 10th Grade On time	Teacher Effect: Behavioral Factor
Teacher Effect: Test Score	1					
Teacher Effect: Suspended	-0.0489	1				
Teacher Effect: Absences	-0.0967	0.159	1			
Teacher Effect: GPA	0.1933	-0.1478	-0.1949	1		
Teacher Effect: In 10th Grade On time	0.1315	-0.1503	-0.0901	0.3616	1	
Teacher Effect: Behavioral Factor	0.164	-0.4606	-0.3448	0.6311	0.6329	1

Notes: This table reports the estimated two-way correlation coefficient between the estimated teacher effects (μ_{zj}) on each outcome and their effects on each other outcome.

Table 5: *Effect of Out of Sample Teacher Effects on 9th Grade Outcomes*

	Outcome: Test Score			Outcome: Behavioral Factor		
	1	2	3	4	5	6
Test Score Effect (sigma)	0.0501**		0.0504**	0.00218		0.00109
	[0.00278]		[0.00280]	[0.00187]		[0.00182]
Behaviors Effect (sigma)		0.0095	-0.00980+		0.0338**	0.0333**
		[0.00710]	[0.00588]		[0.00655]	[0.00661]
Observations	660,434	660,434	660,434	660,191	660,191	660,191

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1
All models include track fixed effects and year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), and the number of honors courses taken during 9th grade.

Table 6: *Effect of Out of Sample Teacher Effects on Longer-Run Outcomes*

	Outcome: Graduate			Outcome: Dropout		
	1	2	3	4	5	6
	OLS	OLS	OLS	OLS	OLS	OLS
Test Score Effect (sigma)	0.00138*		0.00114	-0.000642+		-0.000479
	[0.000695]		[0.000695]	[0.000373]		[0.000373]
Behaviors Effect (sigma)		0.00782**	0.00736**		-0.00517**	-0.00498**
		[0.00251]	[0.00252]		[0.00153]	[0.00154]
% Increase in Variance			249%			527%
Observations	624,078	624,078	624,078	624,078	624,078	624,078

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1
All models include track fixed effects and year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), and the number of honors courses taken during 9th grade.

Table 7: *Effect of Out of Sample Teacher Effects on Longer-Run Outcomes: By Subject*

	1	2	3	4	5	7	8
	9th Grade Outcomes		Longer Run Outcomes				
	Test Score	Behavioral Factor	Dropout	Graduate	Take the SAT	Intentions for 4-Year College	High School GPA (at graduation)
English: Test Score Effect (sigma)	0.0330**	-0.00223	-0.000232	0.000533	-0.000872	-0.000545	-0.00478**
	[0.00274]	[0.00238]	[0.000478]	[0.000858]	[0.00106]	[0.00144]	[0.00138]
Algebra: Test Score Effect (sigma)	0.0720**	0.00503+	-0.00100+	0.00183	0.00279*	0.0022	0.00208
	[0.00491]	[0.00278]	[0.000608]	[0.00113]	[0.00136]	[0.00154]	[0.00172]
English: Behaviors Effect (sigma)	-0.00426	0.0343**	-0.00517**	0.00751**	0.00485+	0.0127**	0.0159**
	[0.00607]	[0.00661]	[0.00155]	[0.00254]	[0.00266]	[0.00415]	[0.00385]
Algebra: Behaviors Effect (sigma)	-0.108	0.0455	0.0372	0.0226	-0.0332	-0.0733	-0.032
	[0.179]	[0.154]	[0.0241]	[0.0437]	[0.0536]	[0.0579]	[0.0717]
Observations	665,382	665,127	624,078	624,078	563,318	563,318	485,099

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1
All models include track fixed effects and year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), and the number of honors courses taken during 9th grade.

Bibliography

1. Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*, 95-135.
2. Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. Science, 333, 1034-1037.
3. Alexander, K. L., Entwisle, D. R., & Thompson, M. S. (1987). School Performance, Status Relations, and the Structure of Sentiment: Bringing the Teacher Back In. *American Sociological Review, 52*, 665-82.
4. Altonji, Joseph G., Todd E. Elder & Christopher R. Taber, 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," Journal of Political Economy, University of Chicago Press, vol. 113(1), pages 151-184, February.
5. Bacher-Hicks, A., Kane, T., & Staiger, D. (2015). Validating Teacher Effect Estimates Using Changes in Teacher Assignment in Los Angeles. Harvard University Working Paper.
6. Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the Big Five in late childhood. *Personality and Individual Differences, 34*(4), 645-664.
7. 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.
8. Booker, K., Sass, T. R., Gill, B., & Zimmer, R. (2011). The Effect of Charter High Schools on Educational Attainment. *Journal of Labor Economics, 29*(2), 377-415.
9. Borghans, L., Weel, B. T., & Weinberg, B. A. (2008). Interpersonal Styles and Labor Market Outcomes. *Journal of Human Resources, 43*(4), 815-58.
10. Bowles, S., Gintis, H., & Osborne, M. (2001). The Determinants of Earnings: A Behavioral Approach. *Behavioral Approach, 39*(4), 1137-76.
11. Brookhart, S. M. (1993). Teachers' Grading Practices: Meaning and Values. *Journal of Educational Measurement, 30*(2), 123-142.
12. Bureau of Labor Statistics Website (2016) http://www.bls.gov/emp/ep_chart_001.htm (Retrieved February 12, 2016)
13. Carneiro, P., Crawford, C., & Goodman, A. (2007). The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes. *CEE Discussion Papers 0092*.
14. Scott E. Carrell & James E. West, 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," Journal of Political Economy, University of Chicago Press, vol. 118(3), pages 409-432, 06.
15. Cascio, E., & Staiger, D. (2012). Knowledge, Tests, and Fadeout in Educational Interventions. *NBER working Paper Number 18038*.
16. Chamberlain, Gary., "Predictive effects of teachers and schools on test scores, college attendance, and earnings" *PNAS 2013 ; October 7, 2013, doi:10.1073/pnas.1315746110*
17. Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. *Quarterly Journal of Economics, 126*(4), 1593-1660.
18. 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): 2593-2632.
19. 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.
20. Deming. (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*(3), 111-134.
21. Deming, D. (2011). Better Schools, Less Crime? *The Quarterly Journal of Economics, 126*(4), 2063-2115.
22. Downey, D., & Shana., P. (2004). When Race Matters: Teachers' Evaluations of Students' Classroom Behavior. *Sociology of Education, 77*, 267-82.
23. Douglass, Harl R.. 1958. "What Is a Good Teacher?". The High School Journal 41 (4). University of North Carolina Press: 110–13.
24. Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101.
25. Duckworth, A. L., & Carlson, S. M. (in press). Self-regulation and school success. In B.W. Sokol, F.M.E. Grouzet, & U. Müller (Eds.), *Self-regulation and autonomy: Social and developmental dimensions of human conduct*. New York: Cambridge University Press.

26. Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? : evidence from NELS88. *Industrial and Labor Relations Review, 48*, 547-561.
27. Figlio, David N. & Morton O. Schapiro & Kevin B. Soter, 2015. "Are Tenure Track Professors Better Teachers?," The Review of Economics and Statistics, MIT Press, vol. 97(4), pages 715-724,
28. Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica, 77*(3), 953-973.
29. Fredriksson, P., Ockert, B., & Oosterbeek, H. (2012). Long-Term Effects of Class Size. *Quartlerly Journal of Economics*.
30. Furnham, A., Monsen, J., & Ahmetoglu, G. (2009). Typical intellectual engagement, Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. *British Journal of Educational Phychology, 79*, 769-782.
31. Grissom, Jason., Loeb Susanna,. and Christopher Doss (2015) "The Multiple Dimensions of Teacher Quality: Does Value-Added Capture Teachers' Nonachievment Contributions to their schools? In "Improving Teacher Evaluation Systems" Edited by Jason Grisson and Peter Youngs.
32. Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job," The Hamilton Project White Paper 2006-01.
33. Greene, William.,2002 "Econometric Analysis" Fifth Edition, Prentice Hall, Upper Saddle River, New Jersey.
34. Harris, D., & Anderson, A. (2012). Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence From Middle School Teachers. *Association of Public Policy Analysis & Management.* Baltimore.
35. Heckman, J. (1999). Policies to Foster Human Capital. *NBER Working Paper 7288*.
36. Heckman, J. J. and T. Kautz (2012, August). Hard evidence on soft skills. Lab. Econ. 19(4), 451–464. Adam Smith Lecture.
37. Heckman, J. J., & Rubinstein, Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review, 91*(2), 145-49.
38. Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics, 24*(3), 411-82.
39. Heckman, J., Pinto, R., & Savelyev, P. (forthcoming). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*.
40. Holmstrom, B., & Milgrom, P. (1991). Multitask Principal-Agent Analysis: Incentive Contracts, Asset Ownership and Job Design. *Journal of Law, Economics and Organization, 7*(Special Issue), 24-52.
41. Howley, A., Kusimo, P. S., & Parrott, L. (2000). Grading and the ethos of effort. *Learning Environments Research, 3*, 229-246.
42. Jacob, Brian, Lefgren, Lars and David Sims. (2010). "The Persistence of Teacher-Induced Learning Gains." Journal of Human Resources 45(4): 915-943.
43. Jackson, C. K. (2014). Teacher Quality at the High-School Level: The Importance of Accounting for Tracks. *Journal of Labor Economics*, Vol. 32, No. 4.
44. Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics,* Volume 95, pp1096-1116.
45. Jackson, C. K., & Bruegmann, E. (2009). Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics, 1*(4), 85-108.
46. Jackson, Kirabo, C., Jonah Rockoff, and Douglas Staiger. "Teacher Effects and Teacher Related Policies" Annual Review of Economics 6.34 (2014).
47. Jencks, C. (1979). *Who Gets Ahead? The Determinants of Economic Success in America.* New York: Basic Books.
48. Jennings, J. L., & DiPrete, T. A. (2010). Teacher Effects on Social and Behavioral Skills in Early Elementary School. *Sociology of Education, 83*(2), 135-159.
49. John, O., Caspi, A., Robins, R., Moffit, T., & Stouthamer-Loeber, M. (1994). The "Little Five": exploring the nomological network of the Five-Factor Model of personality in adolescent boys. *Child Development, 65*, 160–178.
50. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Pearson.
51. Kautz, Tim., Zanoni, Wladimir., "Measuring and Fostering Non-Cognitive Skills in Adolescence: Evidence from Chicago Public Schools and the OneGoal Program." (2014) University of Chicago Mimeo [URL http://home.uchicago.edu/~tkautz/OneGoal_TEXT.pdf]
52. Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental

Evaluation. *NBER working paper 14607*.
53. 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." Seattle, WA: Bill & Melinda Gates Foundation
54. Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-addedmodels. *Quantitative Economics, 3*, 333-362.
55. Koedel, C. (2008). An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Department of Economics, University of Missouri Working Paper 0808*.
56. Koedel, C. (2008). Teacher Quality and Dropout Outcomes in a Large, Urban School District. *Journal of Urban Economics, 64*(3), 560-572.
57. Koedel, C., & Betts, J. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? *Education Finance and Policy, 6*(1), 18-42.
58. Lee, C.D. (2007). *The Role of Culture in Academic Literacies: Conducting Our Blooming in the Midst of the Whirlwind*. Teachers College Press.
59. Lindqvist, E., & Vestman, R. (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-128.
60. Lleras, Christy,. "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 (2008) 888–902.
61. Lounsbury, J. W., Steel, R. P., Loveland, J. M., & Gibson, L. W. (2004). An Investigation of Personality Traits in Relation to Adolescent School Absenteeism. *Journal of Youth and Adolescence, 33*(5), 457–466.
62. Lucas, S. R., & Berends, M. (2002). Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking. *Sociology of Education, 75*(4), 328-348.
63. Mihaly, Kata., Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood (2013). "A Composite Estimator of Effective Teaching" Gates foundation Research Paper.
64. Mansfield, R. (2012). Teacher Quality and Student Inequality. (Working Paper) *Cornell University*.
65. Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," Journal of the American Statistical Association 78(381), 47-55
66. Reardon, S. F., Raudenbush, S. W. (2009). "Assumptions of value-added models for estimating school effects." Education Finance and Policy, 4(4), 492-519.
67. Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73*(2), 417-458.
68. Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*.
69. Sadker, D. M., & Zittleman, K. (2006). *Teachers, Schools and Society: A Brief Introduction to Education.* McGraw-Hill.
70. Siskin Leslie, "Departments As Different Worlds: Subject Subcultures In Secondary Schools" Educational Administration Quarterly, May 1991, pp124-160.
71. Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance."*American Economic Review*, 102(7): 3628-51.
72. Todd, Petra E., and Kenneth I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal* 113 (February): F3–F33.
73. Tyler, John H. and Magnus Lofstrom. 2009. "Finishing High School: Alternative Pathways and Dropout Recovery." Pp. 77-103 in America's High Schools, vol. 19, The Future of Children, Princeton, NJ.
74. Waddell, G. (2006). Labor-Market Consequences of Poor Attitude and Low Self-Esteem in Youth. *Economic Inquiry, 44*(1), 69-97.

Appendix

Appendix 1: *Matching Teachers to Students*

The teacher ID in the testing file corresponds to the teacher who administered the exam, who is not always the teacher that taught the class (although in many cases it will be). To obtain high-quality student-teacher links, I link classrooms in the End of Course (EOC) testing data with classrooms in the Student Activity Report (SAR) files (in which teacher links are correct). The NCERDC data contains End of Course (EOC) files with test-score-level observations for a certain subject in a certain year. Each observation contains various student characteristics, including ethnicity, gender, and grade level. It also contains the class period, course type, subject code, test date, school code, and a teacher ID code. Following Mansfield (2012), I group students into classrooms based on the unique combination of class period, course type, subject code, test date, school code, and the teacher ID code. I then compute classroom-level totals for 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 or English I teacher are automatically linked to the teacher ID from the SAR files. These are perfectly matched. Matched classes are set aside.
(2) Classes that match exactly on all classroom characteristics and the teacher ID are deemed matches. These are deemed perfectly matched. Matched classes are set aside.
(3) Compute a score for each potential match (the sum of the squared difference between each observed classroom characteristics for classrooms in the same school in the same year in the same subject, and infinity otherwise) in the SAR file and the EOC data. Find the best match in the SAR file for each EOC classroom. If the best match also matches in the teacher ID, a match is made. These are deemed imperfectly matched. Matched classes are set aside.
(4) Find the best match (based on the score) in the SAR file for each EOC classroom. If the SAR classroom is also the best match in the EOC classroom for the SAR class, a match is made. These are deemed imperfectly matched. Matched classes are set aside.
(5) Repeat step 4 until no more-high quality matches can be made.

This procedure leads to a matching of approximately 75 percent of classrooms. Results are similar when using cases when the matching is exact, so error due to the fuzzy matching algorithm does not generate any of the empirical findings.

Appendix 2: *Correlations Between Short Run Outcomes*

The correlations among the 9th grade outcomes 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≈0.6) 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 -0.156 for Algebra test scores and -0.097 for English test scores, and the correlations between being suspended are about -0.13 for both Algebra and English test scores. While slightly higher, the correlation between on-time progression to 10th grade (i.e. being a 10th grader the following year) and test scores is only 0.29. 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, Table 2 indicates that test scores predict less than five percent of the variability in absences and being suspended, less than 10 percent of the variability in on-time grade progression, and just over one-third of the variability 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 behavioral outcomes 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 the other outcomes. Similarly, the correlations between absences and test scores are smaller than those between absences and the other outcomes. The third notable pattern is that GPA is relatively well correlated with both the test score and the non-test score outcomes. The fact that GPA is correlated with both test scores and non-test-score outcomes is consistent with research (e.g., Howley, Kusimo, & 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 non-cognitive skills.

In sum, the patterns suggest that the outcomes can be put into three categories; academic aptitude variables (English I and Algebra I 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 these three groups of variables may reflect a somewhat different combination of cognitive and non-cognitive skills. If teachers improve student outcomes through improving both cognitive and non-cognitive skills, their effect on a combination of these outcomes should better predict their effect on longer-run outcomes than using their effects on test scores alone.

Appendix 2 Table 1: *Raw two-way correlation coefficients between outcomes (537,241 Observations)*

	Log of # Days Absent	Suspended	Grade Point Average	In 10th grade on time	Algebra Score 9th Grade	English Score 9th Grade	Behavioral Factor	Test Score Factor
Ln of # Days Absent	1							
Suspended	0.191	1						
Grade Point Average	-0.276	-0.194	1					
In 10th grade on time	-0.181	-0.151	0.447	1				
Algebra Score 9th Grade	-0.156	-0.128	0.59	0.294	1			
English Score 9th Grade	-0.097	-0.127	0.531	0.29	0.618	1		
Behavioral Factor							1	
Test Score Factor							0.5324	1

The behavioral factor was uncovered using factor analysis and is a linear combination of all the non-test score short-run outcomes. Specifically, this non-cognitive factor is 0.64*(GPA)+0.57*(in 10^{th} grade)-0.33*(suspended)-0.45*(log of 1+absences). The weighted average is then standardized to be mean zero, unit variance. The test score factor is the equal weight average of the test score outcomes. It is also standardized to be unit variance and mean zero.

Appendix 3: *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. Appendix 3 Table 1 presents the same models using the NELS-88 data. I predict longer run outcomes as a function of the same behavioral outcomes and test score variables as used in the NCERDC data. All models control for ethnicity, gender, family income, family size, and school fixed effects. The results are consistent with those from the NCERDC data. For both dropout and high-school graduation, increases in the behavioral factor are associated with large effects on longer-run outcomes conditional on test scores. Looking at college going, a 1σ increase in the test score factor (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 behavioral factor is associated with a 9.5 percentage point increase.

The NELS-88 data also include longer-run outcomes from when the respondent was 25 years old. These allow one to see how this behavioral factor (based on 8th grade outcomes) predicts being arrested (or having a close friend who was arrested), employment, and labor market earnings, conditional on 8th grade test scores. The results show that test scores are actually positive associated with being arrested (conditional on all the covariates), but a 1σ increase in the behavioral factor 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 behavioral factor 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 behavioral factor is associated with a similar 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 behavioral factor is associated with 24.6 percent higher earnings.

In recent findings, both Lindqvist & Vestman (2011) and Heckman, Stixrud, & Urzua (2006) find that non-cognitive ability is particularly important at the lower end of the earnings distribution. Insofar as the behavioral factor truly captures non-cognitive skills, one would expect this to be the case for this factor also. To test for this, I estimate quantile regressions to obtain the marginal effect on log wages at different points in the earnings distribution. The results (appendix table A4) show that at the 90th percentile through the 75th percentile of the earnings distribution, a 1σ increase in test scores and the behavioral factor is associated with a very similar increase of between 5 and 6 percent higher earnings. However, at the median level the behavioral factor is more important; the marginal effect of a 1σ increase in test scores and the behavioral factor is 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 behavioral factor is associated with 23 percent higher earnings. These findings are remarkably similar to those presented in Lindqvist & Vestman (2011) using phychometric measurs of noncogntive skills, suggesting that this factor is a reasonable proxy for non-cognitive ability.

Appendix 3 Table 1: *Relationship Between Short-run Outcome 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 factor: 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]
Behavioral factor: 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

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

All models control for ethnicity, gender, family income, family size, and school fixed effects.

Appendix 3 Table 2: *Effect of test scores and the behavioral factor in 8th grade on adult earnings at different percentiles (NELS-88)*

	Natural log of Income (age 25): Conditional of Working			
Percentile	25th	50th	75th	90th
Test Score factor: z-score	0.00312	0.0318***	0.0495***	0.0582***
	[0.0511]	[0.00939]	[0.00691]	[0.00866]
Behavioral factor: z-score	0.233***	0.100***	0.0679***	0.0509***
	[0.0467]	[0.00858]	[0.00632]	[0.00791]
Observations	10,792	10,792	10,792	10,792

Standard errors in brackets

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

All models control for ethnicity, gender, family income, family size, and school fixed effects.

Appendix 4: *Formal proofs of claims in Sections II and IV.*

Claim: *Teacher effects on y_2 will increase the explained teacher-level variability in the long-run outcome iff $cor(f(\omega_{cj}), g(\omega_{cj})) \neq 0$.*

Proof: The variability in the long run effect explained by the effect on test scores (in a linear regression model) is simply $A \equiv var(\gamma_{1a}\theta_{1j})$, where γ_{1a} is the coefficient on θ_{1j} in a simple linear regression predicting θ_{lj}. In a model with both the effect on test scores and the effect on outcome 2, the explained variance is $B \equiv var(\gamma_{1b}\theta_{1j} + \gamma_{2b}\theta_{2j})$, where γ_{1b} and γ_{2b} are the coefficient on θ_{1j} and θ_{2j} in a multivariable linear regression predicting θ_{lj}, respectively.

From Green (2002), $B \equiv var(\gamma_{1b}\theta_{1j} + \gamma_{2b}\theta_{2j}) = var(\gamma_{1a}\theta_{1j} + \gamma_{2a}\ddot{\theta}_{2j})$ 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 effect on long 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 $var(\ddot{\theta}_{2j}) > 0$, the explained variance will be greater with effects on both outcomes than with only test score value-added (i.e. B>A) if $\gamma_{2a} \neq 0$. Because $\gamma_{2a} = cov(f(\omega_{cj}), g(\omega_{cj}))/var(g(\omega_{cj}))$, it follows that $\gamma_{2a} = 0$ if $cov(f(\omega_{cj}), g(\omega_{cj})) = 0$.

Claim: *With two dimensions of ability including lagged values of two linearly independent outcomes is sufficient to proxy for students' incoming skills in both the cognitive and non-cognitive dimensions.*

Proof: Take two linearly independent outcomes 1 and 2 such that $y_1 = \upsilon_{ci}\beta_{1c} + \upsilon_{ni}\beta_{1n}$ and $y_2 = \upsilon_{ci}\beta_{2c} + \upsilon_{ni}\beta_{2n}$. It follows that $\upsilon_{ni} = \left(\frac{y_1}{\beta_{1c}} - \frac{y_2}{\beta_{2c}}\right)/\left(\frac{\beta_{1n}}{\beta_{1c}} - \frac{\beta_{2n}}{\beta_{2c}}\right) = y_1/\left(\beta_{1c}\left(\frac{\beta_{1n}}{\beta_{1c}} - \frac{\beta_{2n}}{\beta_{2c}}\right)\right) + y_2/\left(\beta_{2c}\left(\frac{\beta_{1n}}{\beta_{1c}} - \frac{\beta_{2n}}{\beta_{2c}}\right)\right)$ and that $\upsilon_{ci} = y_1/\left(\beta_{1n}\left(\frac{\beta_{1c}}{\beta_{1n}} - \frac{\beta_{2c}}{\beta_{2n}}\right)\right) + y_2/\left(\beta_{2n}\left(\frac{\beta_{1c}}{\beta_{1n}} - \frac{\beta_{2c}}{\beta_{2n}}\right)\right)$. Because υ_{ni} and υ_{ci} are linear functions of the two outcomes, they are proxies for υ_{ni} and υ_{ci} (Green 2002). It follows that a linear regression that conditions on *any* two linearly independent outcomes will yield the same coefficient on all *other* covariates as a regression model that included direct measures of cognitive and noncognitive skills prior to high school entry.[27]

[27] This argument abstracts away from problems associated with measurement error in outcomes 1 and 2. With such measurement error, the two outcomes may serve as imperfect proxies and including additional short-run outcomes should mitigate this problem.

Appendix 5: *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 9th 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 Appendix 5 Table 1. As highlighted in Jackson (2014) and Harris and Anderson (2012), it is not only the course that matters but also the levels at which students take a course. As such, following Jackson (2014), a school track is the unique combination of the ten largest 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/course-group/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. Because many students pursue the same course of study, less than one percent of all students are in singleton tracks, 80 percent of students are in tracks with more than 30 students, and the average student is in a school track with 179 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 computed the number of teachers in each non-singleton school-track-year-cell for both Algebra I and English I (Appendix 5 Table 2). About 63 and 51 percent of all school-track-year cells include one teacher in English I and Algebra I, respectively. As such, much variation is likely based on comparing single teachers across cohorts within the same school track. Appendix 7 shows that results using variation within school-track-cohort cells are similar to those obtained using only variation entire across cohorts within a school.

Appendix 5 Table 1: *Most common academic courses*

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

Appendix 5 Table 2: *Distribution of Number of Teachers in Each School-Track-Year Cell*

	Percent	
Number of Teachers in School-Track-Year Cell	English	Algebra
1	63.37	51.07
2	18.89	26.53
3	9.12	11.00
4	5.60	6.38
5	3.03	3.25
6	0	1.77

Note: This is after removing singleton tracks.

Appendix 6: *Showing Effect of Teachers on Individual Behavioral Outcomes*

To show that the relationship between longer-run outcomes and teacher effects on the behavioral factor are not driven by any single behavior, I estimate equation [7] where instead of using the teacher effects on the behavioral factor that combines all behaviors in a single variable, I use the teacher effect on the individual behaviors separately. Because many of the outcomes are binary, the Empirical Bayes approach does not increase precision and in many cases reduces it. As such, I present results using the unadjusted leave-year-out teacher-level mean residuals ($\hat{\theta}_{zj,-t}$). In addition to presenting results for teacher effects on each behavioral outcome, I also present results using a behavioral factor that is based only on absences, suspensions, and on-time grade progression (that is, excluding GPA).

For both graduation and dropout (Appendix 6 Tables 1 and 2), the teachers effect on the factor excluding GPA predict the longer run outcomes—showing that the GPA variable does not drive the results. One can also see that teacher effects on suspensions, GPA, and on time grade progression each independently predict teacher effect on the longer run outcomes —showing that no single variable drives the results. Finally, teacher effects on the behavioral factor that combines all the behaviors is more strongly associated with improved longer run outcome than the effect in each of the individual outcomes – indicating that it is improvement in those skills common to *all the behaviors* that is driving the results.

Appendix 6 Table 1: *Effects of Individual Teacher Effects on Longer Run Outcomes*

	Outcome: Graduate High School					
	1	2	3	4	5	6
Effect: Test Score	0.000611	0.00067	0.000941	0.000914	0.000636	0.000909
	[0.00122]	[0.00121]	[0.00122]	[0.00121]	[0.00123]	[0.00120]
Effect: Behavioral Factor	**0.00743***					
	[0.00351]					
Effect: Behavioral Factor w/o GPA		**0.00874+**				
		[0.00449]				
Effect: Suspended			**-0.0374+**			
			[0.0199]			
Effect: Absences				-0.00116		
				[0.000777]		
Effect: GPA					0.00468	
					[0.00306]	
Effect: In 10th on time						0.00333
						[0.00455]
Observations	621,259	621,259	621,259	621,259	621,259	621,259

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

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

All models include track fixed effects and year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), and the number of honors courses taken during 9th grade.

Appendix 6 Table 2: *Effects of Individual Teacher Effects on Longer Run Outcomes*

	Outcome: Graduate High School					
	1	2	3	4	5	6
Effect: Test Score	-0.00031	-0.00044	-0.00053	-0.00059	-0.00031	-0.00036
	[0.000627]	[0.000623]	[0.000626]	[0.000630]	[0.000636]	[0.000626]
Effect: Behavioral Factor	**-0.00436***					
	[0.00214]					
Effect: Behavioral Factor w/o GPA		**-0.00351+**				
		[0.00216]				
Effect: Suspended			0.0127			
			[0.0119]			
Effect: Absences				-7.65E-05		
				[0.000434]		
Effect: GPA					**-0.00283+**	
					[0.00149]	
Effect: In 10th on time						**-0.00450+**
						[0.00231]
Observations	621,259	621,259	621,259	621,259	621,259	621,259

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

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

All models include track fixed effects and year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), and the number of honors courses taken during 9th grade.

Appendix 7: *Addressing Selection*

The key identifying assumption is that conditional on controls for tracking and 8th grade outcomes, there is no selection of students to teachers. To present evidence that the results are not driven by selection, I present a test of selection on observables following Chetty et al (2014b). Specifically, I predict each outcome (based on a linear regression of each outcome on 7th grade math and reading scores, parental education, gender, and ethnicity). I then estimate equation [7] on the predicted outcomes while excluding parental education, gender, ethnicity, and 7th grade math and reading scores from the set of covariates. To exploit only the variation within tracks, the models include track-year fixed effects. If the estimated effects were driven by positive selection to teachers on observables, one might observe a positive relationship between the estimated teacher effects and the predicted outcomes. Results are in Appendix 7 Table 1. Columns 3 and 4 show that the estimated teacher are unrelated to predicted outcomes (i.e. unrelated to parental education, gender, ethnicity, and 7th grade test score which are all strong predictors of the longer run outcomes), so that there is no selection on observables.

To test for selection on unobservables within school track cohorts, I follow Chetty, Friedman, and Rockoff (2014a) 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 the estimated teacher effects merely capture student selection to teachers within school cohorts, then the arrival of a teacher with a high positive predicted effect (who increases the average predicted teacher effect 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 predicted effects are real, differences in average predicted teacher quality across cohorts (driven by changes in teaching personnel within schools over time) should be associated with similar differences across cohorts in average cohort-level outcomes as the same difference in estimated teacher quality across individual students within the same cohort.

An intuitively appealing test of this would be to aggregate the treatment to the school-year level and determine whether changes in the school average teacher quality lead to the same effects as individual changes in teacher quality. This is the test implemented in Chetty et. al. (2014a). Result of such a test are in Appendix 7 Table 1. Columns 5 and 6 report the marginal effect of the school mean estimated teacher effects on test scores and on the behavioral factor (in models that include school fixed effects and year fixed effects only). For both of these outcomes, one can reject the null hypothesis that the average teacher effect on behaviors is zero. This shows that the positive effects observed within tracks were not driven by selection on unobservable. For both outcomes, the effects of the average teacher effects are larger than those at the individual teacher level within tracks (Columns 1 and 2). However, one cannot reject the null hypothesis that they are the same.

Even though one cannot reject the hypothesis that the individual level variation and the cohort level variation are the same. It is worth exploring why the point estimates are larger using the aggregate variation. A likely explanation is measurement error. Due to measurement error, the two approaches could yield different results even if there is no selection. Because aggregation reduces the variability of the measurement error, the signal to noise ratio may be higher in the school level average than in the individual teacher effects. Accordingly, one would expect that the coefficient on mean teacher quality will tend to be larger than that on individual teacher quality. This is what one observes. To allow for an apples-to-apples comparison, I also propose instrumental variables specifications in the spirit of Chetty at al (2014b), that are robust to differences in the noisiness of the school-level means versus the individual teacher estimates. A way to test for whether teacher induced changes in 9th grade outcomes are driven by selection within a cohort is to estimate and compare two instrumental variables regressions that rely on completely distinct sources of variation. Because this approach uses the teacher quality

estimates (and the school year averages of these estimates) as instruments, differences in the signal-to-noise ratio between the set of instruments will not affect the 2SLS coefficient, and will be reflected in the standard errors.

Model 1: Regress the longer-run outcomes on 9th grade test scores, the 9th grade behavioral factor, covariates, and track-school-year fixed effects. One can then instrument for 9th grade test scores and the behavioral factor in 9th grade with the estimated individual teacher effects. Because this model includes track-school-year fixed effects, it compares the outcomes of individual students who had higher test scores or behavior factor in 9th grade *because they were exposed to teachers with different estimated effects within the same track and school and year*. This instrumental variables model uses only the variation in 9th grade outcomes driven by the potentially endogenous variation in predicted teacher effects across teacher within cohorts of 9th graders at a given school in a given track.

Model 2: Regress the longer-run outcomes on 9th grade test scores, the 9th grade behavioral factor, covariates, and separate school fixed effects and year fixed effect. One can then instrument for 9th grade test scores and the behavioral factor in 9th grade with the average estimated individual teacher effects across all 9th graders in that school in that year. Because this model includes separate school fixed-effects and year fixed-effects and aggregates the treatment to the school-year level, it compares the outcomes of all 9th graders in a given cohort at a given school to those of other entre cohorts within the same school but who were exposed to different levels of average estimated teacher effects due to changes in the personnel at the school over time. This model excludes the potentially endogenous variation in estimated teacher quality at the individual teacher level that could occur within school cohorts (exploited in model 1) and u*ses only the arguably selection-free variation in teacher quality across school years*.

If the two distinct sources of variation yield similar 2SLS coefficients on 9th grade test scores and the behavioral factor, it would be compelling evidence that the estimated effects are real and are not driven by selection to teachers within a given 9th grade cohort at a given school. Appendix 7 Table 2 presents the estimated 2SLS regressions of 9th grade outcomes on high-school graduation and dropout. *Note that the treatment variable is the 9th grade outcomes using the teacher effects as instruments.* For both outcomes, results using the clean variation across cohorts yield almost identical point estimates as those based on the potentially endogenous variation within cohorts. Looking to graduation (columns 1 and 2), the 2SLS coefficient on test scores is 0.022 (p-value>0.1) using only the within school cohort variation and 0.055 (p-value>0.1) using only the average across cohort variation driven by personnel changes within schools over time. One cannot reject the hypothesis of equality across the two models. Similarly, the 2SLS coefficient on the behavioral factor is 0.209 (p-value<0.01) using only the within school cohort variation and 0.135 (p-value<0.05) using only the average across cohort variation driven by personnel changes within schools over time. Again, one cannot reject the hypothesis of equality across the two models. For dropout (columns 5 and 6) the coefficient estimates are also very similar using both distinct sources of variation, and one cannot reject the hypothesis of equality across the two models. Consistent with other studies that seek to validate teacher effects in value-added models (e.g. Chetty et al (2014b), Kane and Stagier (2008), Kane et al (2013) and Backer Hicks et al 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.

Appendix 7 Table 1: *Effect on Predicted Longer Run Outcomes*

	Outcome: Graduation	Outcome: Dropout	Outcome: Predicted Graduation	Outcome: Predicted Dropout	Outcome: Graduation	Outcome: Dropout	Outcome: Predicted Graduation	Outcome: Predicted Dropout
	1	2	3	4	5	6	7	8
Test Score Effect (sigma)	0.00112	-0.000555	0.000159	-0.000025				
	[0.000682]	[0.000370]	[0.000110]	[0.000033]				
Behaviors Effect (sigma)	0.00600*	-0.00345*	-0.000612	0.000126				
	[0.00234]	[0.00146]	[0.000393]	[0.000113]				
Mean [Test Score Effect (sigma)]					0.00623	0.000963	0.000339	0.000048
					[0.00465]	[0.00210]	[0.00129]	[0.000458]
Mean [Behaviors Effect (sigma)]					0.0311*	-0.0308**	0.000694	-0.000355
					[0.0155]	[0.00847]	[0.00354]	[0.00121]
Observations	660,434	660,191	660,434	660,191	660,434	660,191	660,434	660,191

Robust standard errors in brackets

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

Notes: Specifications 1,2,3 and 4 exploit teacher-level variation within school tracks. These models include school-track-year fixed effects , 8th grade outcomes (math and reading scores, repeater status, ever suspended, and attendance), classroom averages of these 8th grade outcomes, and the number of honors courses taken during 9th grade. Standard errors are adjusted for clustering at the track level in these models. Predicted outcomes are fitted valued of a regression predicting each outcome as a function of on 7th grade math and reading scores, parental education, gender, and ethnicity. Specifications 5,6,7 and 8 exploit school-cohort level variation within schools (across tracks). These models include school fixed effects and year fixed effects only. Standard errors in these models are adjusted for clustering at the school level.

Appendix 7 Table 2: *Testing for Selection on Unobserables*

	Graduate		Predicted Graduate		Dropout		Predicted Dropout	
	1	2	3	4	5	6	7	8
	2SLS: Individual Teacher Effects	2SLS: School mean of Teacher Effects	2SLS: Individual Teacher Effects	2SLS: School mean of Teacher Effects	2SLS: Individual Teacher Effects	2SLS: School mean of Teacher Effects	2SLS: Individual Teacher Effects	2SLS: School mean of Teacher Effects
Test Score	0.0222	0.0551	-0.00088	-0.00239	-0.00102	0.0498	-0.00088	-0.00239
	[0.0179]	[0.0652]	[0.00126]	[0.00519]	[0.0100]	[0.0453]	[0.00126]	[0.00519]
Behavioral Factor	0.209**	0.135*	0.00266	0.0048	-0.138**	-0.171**	0.00266	0.0048
	[0.0502]	[0.0679]	[0.00319]	[0.00449]	[0.0342]	[0.0528]	[0.00319]	[0.00449]
Observations	623,827	687,040	662,018	730,933	623,827	687,040	662,018	730,933

Robust standard errors in brackets

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

Models that exploit individual teacher-level variation in predicted effects (Columns 1,3,5, and 7) include track-school-year fixed effects. Models that exploit average school cohort-level variation in average predicted effects (Columns 1,3,5, and 7) include school fixed effect and year fixed effects.

Models that predict actual outcomes (columns 1,2,5 and 6) include track fixed effects and year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), the number of honors courses taken during 9th grade, and indicator variables for each track.

Models that predict predicted outcomes (columns 3,4,7 and 8) include 8th grade outcomes (math and reading scores, repeater status, ever suspended, and attendance), classroom averages of these 8th grade outcomes, the number of honors courses taken during 9th grade, and indicator variables for each track. Predicted outcomes are fitted valued of a regression predicting each outcome as a function of on 7th grade math and reading scores, parental education, gender, and ethnicity.

Appendix 8: *Additional Tables*

Appendix 8 Table 1: *Effect of Out of Sample Teacher Effects on Longer-Run Outcomes with School-by-Year Fixed Effects*

	Outcome: Graduate			Outcome: Dropout		
	1	2	3	4	5	6
	OLS	OLS	OLS	OLS	OLS	OLS
Test Score Effect (sigma)	0.000876		0.000671	-0.00055		-0.00044
	[0.000648]		[0.000648]	[0.000359]		[0.000361]
Behaviors Effect (sigma)		0.00674**	0.00647**		-0.00370*	-0.00352*
		[0.00226]	[0.00227]		[0.00146]	[0.00147]
Observations	624,078	624,078	624,078	624,078	624,078	624,078

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1
All models include track fixed effects and school-by-year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), and the number of honors courses taken during 9th grade.

Appendix 8 Table 2: *Effect of Out of Sample Teacher Effects on Longer-Run Outcomes controlling for Both 8th and 7th grade Behaviors*

	Outcome: Graduate			Outcome: Dropout		
	1	2	3	4	5	6
	OLS	OLS	OLS	OLS	OLS	OLS
Test Score Effect (sigma)	0.00148*		0.00123+	-0.000621+		-0.00047
	[0.000691]		[0.000692]	[0.000373]		[0.000373]
Behaviors Effect (sigma)		0.00821**	0.00771**		-0.00495**	-0.00476**
		[0.00249]	[0.00251]		[0.00151]	[0.00152]
Observations	592,954	592,954	592,954	592,954	592,954	592,954

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1
All models include track fixed effects and school fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th and 7th grade, ever suspended in 8th and 7th grade, and attendance in 8th and 7th grade), classroom averages of these 8th grade outcomes, student-level demographics (parental education, ethnicity, and gender), and the number of honors courses taken during 9th grade.

Appendix 8 Table 3: *Observable Teacher Correlates of the Behavioral Factor*

	1	2	3	4	5	6	7	8
	English and Algebra Teachers				English Teachers Only			
	Test Scores	Behavioral	Graduate	Dropout	Test Scores	Behavioral	Graduate	Dropout
Racial Match	0.00398	0.00512	0.00540**	-0.000328	0.00195	0.0046	0.00423	0.000478
	[0.00404]	[0.00481]	[0.00207]	[0.00101]	[0.00352]	[0.00610]	[0.00269]	[0.00130]
Gender Match	0.0397**	0.00359	9.93E-05	-0.000464	0.00712	0.00263	-0.00161	0.00115
	[0.00583]	[0.00522]	[0.00244]	[0.00126]	[0.00483]	[0.00712]	[0.00319]	[0.00164]
Ln(Years of Experience)	0.00192	-0.00227	-0.00131	0.000229	0.00596**	-0.00555+	-0.00183+	0.000362
	[0.00280]	[0.00218]	[0.000810]	[0.000444]	[0.00207]	[0.00293]	[0.00104]	[0.000596]
Certified	0.0164*	0.00331	0.00493+	-0.00117	0.0112+	0.00568	0.00788*	-0.00336+
	[0.00772]	[0.00777]	[0.00274]	[0.00150]	[0.00597]	[0.00949]	[0.00352]	[0.00195]
Average Test Score	0.00126	-0.00125	2.22E-05	-0.000182	0.00284	0.00147	0.00129	-0.000672
	[0.00324]	[0.00223]	[0.000829]	[0.000454]	[0.00247]	[0.00319]	[0.00120]	[0.000679]
Advanced Degree	-0.00492	0.00610+	0.000359	-0.000605	0.000772	0.0106*	0.00174	-0.000888
	[0.00470]	[0.00342]	[0.00130]	[0.000666]	[0.00342]	[0.00430]	[0.00174]	[0.000915]
75th%ile SAT at College	0.000104*	-0.000045	-2.21E-06	5.34E-08	0.000033	-0.000086*	-3.21E-06	3.37E-06
	[0.000042]	[0.000029]	[1.11e-05]	[6.01e-06]	[0.000029]	[0.000039]	[1.41e-05]	[8.19e-06]
Fully Licensed	0.0182**	-0.00219	0.00432*	-0.000053	0.00128	-0.00408	0.00575*	-0.000133
	[0.00642]	[0.00573]	[0.00217]	[0.00117]	[0.00484]	[0.00750]	[0.00281]	[0.00152]
Licensed in Math	0.0516**	0.00925	-0.00379	7.42E-05				
	[0.0157]	[0.0126]	[0.00690]	[0.00296]				
Observations	566,090	565,884	566,090	566,090	378,575	378,419	378,575	378,575

Robust standard errors in brackets
** p<0.01, * p<0.05, + p<0.1

All models include track fixed effects and year fixed effects, incoming outcomes (math and reading scores in both 7th and 8th grades, repeater status in 8th grade, ever suspended in 8th grade, and attendance in 8th grade), classroom averages of these lagged outcomes, student-level demographics (parental education, ethnicity, and gender), the number of honors courses taken during 9th grade, and indicator variables for each track.