2018

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

C. Kirabo Jackson

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

C. Kirabo Jackson

Northwestern University and National Bureau of Economic Research

Teachers affect a variety of student outcomes through their influence on both cognitive and noncognitive skill. I proxy for students' noncognitive skill using non–test score behaviors. These behaviors include absences, suspensions, course grades, and grade repetition in ninth grade. Teacher effects on test scores and those on behaviors are weakly correlated. Teacher effects on behaviors predict larger impacts on high school completion and other longer-run outcomes than their effects on test scores. Relative to using only test score measures, using effects on both test score and noncognitive measures more than doubles the variance of predictable teacher impacts on longer-run outcomes.

I. Introduction

At the broadest level, a good teacher is one who teaches students the skills needed to be productive adults (Douglass 1958; Jackson, Rockoff, and Staiger 2014). Not every skill needed in adulthood is well captured by performance on achievement tests. Indeed, a large body of research demonstrates that "noncognitive" skills not captured by standardized tests, such as adaptability, self-restraint, and motivation, are key determinants of adult

I thank David Figlio, Jon Guryan, Simone Ispa-Landa, Clement Jackson, Shayna Silverstein, Mike Lovenheim, James Pustejovsky, Jonah Rockoff, Dave Deming, Jim Heckman, Alexey Makarin, Laia Navarro-Sola, and four anonymous reviewers for insightful comments and feedback. I also thank Kara Bonneau from the North Caroline Education Research Data Center (NCERDC). This publication was made possible in part by the Smith Richardson Foundation and the Carnegie Corporation on New York. The statements made and views expressed are solely the responsibility of the author. Data are provided as supplementary material online.

Electronically published August 30, 2018
[*Journal of Political Economy*, 2018, vol. 126, no. 5]

outcomes.[1] Even so, economists have focused on test score measures of teacher quality (referred to as test score value-added) because they are often the best available measure of student skills.[2] However, good teachers may affect students much more broadly than through their impact on achievement test scores.

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

This paper estimates teacher effects on both test scores and measures of noncognitive skills.[3] I refer to a teacher's effect on a skill measure as value-added. I demonstrate that teachers have meaningful value-added on both test scores and noncognitive skill measures in ninth grade. Surprisingly, test score and noncognitive value-added for the same teacher are weakly correlated ($r = .15$). I show that ninth-grade teachers who raise students' noncognitive skill measures have important impacts on students' life chances. A one standard deviation increase in noncognitive value-added increases students' likelihood of graduating from high school by 1.47 percentage points, compared to only 0.12 percentage points for test score value-added. This pattern of larger impacts of noncognitive value-added replicates across several high school outcomes, including grade progression, SAT taking, twelfth-grade grade point average (GPA), and intentions to attend a 4-year college.

To motivate the empirical work, I follow Cunha and Heckman (2008) and extend the standard test score value-added model that assumes unidimensional student ability. In this extended model, student outcomes

[1] See Heckman and Rubinstein (2001), Waddell (2006), Borghans, Weel, and Weinberg (2008), and Lindqvist and Vestman (2011). Consistent with this, some interventions that have no effect on test scores have meaningful effects on long-term outcomes (Deming 2009, 2011; Booker et al. 2011), and improved noncognitive skills explain the effect of some interventions (Fredriksson, Ockert, and Osterbeek 2013; Heckman, Pinto, and Savelyev 2013).

[2] Having a teacher at the 85th vs. the 15th percentile of the test score value-added distribution is found to increase test scores by between 8 and 20 percentile points (Rivkin, Hanushek, and Kain 2005; Kane and Staiger 2008).

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

are a function of both cognitive and noncognitive skills (Heckman, Stixrud, and Urzua 2006). I propose that one can measure teacher value-added on multiple ninth-grade skill measures but focus on two: test scores and some other outcome. I show that, as long as the two skill measures do not reflect the same exact mix of student abilities, one can better predict teacher impacts on longer-run student outcomes using value-added on both skill measures than using test score value-added alone. I then test this implication empirically.

I employ administrative data on all public school ninth graders in North Carolina from 2005 through 2012. These data contain student scores on math and English exams linked to their subject teachers. To obtain measures of student skills in ninth grade that may not be well captured by test scores, I follow a literature that uses behaviors as proxies for noncognitive skills.[4] To summarize these behaviors with a single variable, I use principal component analysis to create a weighted average of grades, on-time grade progression, absences, and suspensions. I refer to this weighted average of ninth-grade behaviors as the "behavior index." I use the behavior index to measure noncognitive skills that may be missed by test scores. However, I do not claim that this index is unrelated to cognitive skills, nor is this index a comprehensive measure of all noncognitive skills. Using value-added models, I estimate ninth-grade teacher effects on both test scores and behaviors. I then examine how teachers with high noncognitive value-added (i.e., those who improve behaviors) influence longer-run outcomes such as high school completion, SAT taking, and intended college going in ways unmeasured by their test score value-added.[5]

Teachers in ninth grade have meaningful effects on both test scores and those behaviors that proxy for noncognitive skills. Teacher value-added on test scores and the behaviors are weakly correlated ($r = .15$), and, conditional on test score value-added, there is considerable variability in behaviors value-added. In models that predict high school graduation using only test score value-added, a one standard deviation increase in test score value-added raises the likelihood of high school graduation by 0.15 percentage points. However, when also including behaviors value-added, a one standard deviation increase in test score value-added leads to a 0.12 higher likelihood of graduation, and a one standard deviation

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

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

increase in behaviors value-added leads to a 1.47 percentage point higher likelihood of graduating from high school. These results suggest that (*a*) many teachers who raise test scores do not improve behaviors and vice versa, and (*b*) behaviors value-added detects effects on important skills that are not detected by test scores. Including both value-added measures more than doubles the predictable teacher-level variability in high school graduation. Patterns are similar for dropout, SAT taking, and college plans.

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

The results support an idea that many believe to be true but that has not previously been shown: that teacher effects on test scores capture only a fraction of teacher effects on human capital. This underscores the need for evaluations that account for effects on both cognitive and noncognitive skills (Heckman 1999). Because some of the non–test score outcomes used can be manipulated by teachers, using them directly for accountability or evaluation purposes is unwise. However, I present some feasible policy uses. The results provide an explanation for why Chamberlain (2013) finds that test score value-added may reflect less than one-fifth of the total effect of teachers. Also, consistent with Heckman et al. (2013), teacher effects on proxies for noncognitive skills offer an explanation for why teacher test score effects fade over time (Jacob, Lefgren, and Sims 2010) despite having meaningful effects on long-run outcomes.

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

II. Data and Relationships between Variables

I seek to obtain estimates of the effect of ninth-grade teachers on both test scores and proxies for noncognitive skills in ninth grade. I will then explore whether these estimated effects on ninth-grade skill measures predict teacher impacts on longer-run outcomes. I use data on all public school ninth-grade students in North Carolina between 2005 and 2012 obtained from the North Carolina Education Research Data Center. The data include demographics, transcript data, test scores in grades 7–9,

and codes linking student test scores to the teacher who administered the test.[6] I focus on students who took English (English I) and math (algebra I, geometry, or algebra II) courses during ninth grade. Roughly 93 percent of all ninth graders take both English I and one of these math courses. To avoid any bias that would result from teachers influencing students' ninth-grade repetition, I use only the first observation of ninth-grade repeaters.[7] Summary statistics are presented in table 1.

These data cover 573,963 ninth-grade students in 872 secondary schools, with 5,195 English teachers and 6,854 math teachers. The gender split is roughly even. The sample is 58.8 percent white, 26.1 percent black, 7.2 percent Hispanic, and 2.1 percent Asian. Regarding the highest level of education obtained by either of the student's two parents, 46 percent had a high school diploma or less, 14.9 percent had a junior college or trade school degree, 29.4 percent had a 4-year college degree or higher, and 9.5 percent are missing data on parental education. All test score variables are standardized to be mean zero, unit variance, for the full population taking each test during each testing year. Test scores in the sample are higher than average because the ninth graders successfully matched to their classroom teacher are slightly higher-achieving on average.[8]

Informed by studies that use behaviors as proxies for noncognitive skills not measured well by test scores (Lleras 2008; Bertrand and Pan 2013; Kautz and Zanoni 2014; Heckman et al. 2016), I proxy for noncognitive skills using non–test score behaviors available in the data: the log of the number of absences in ninth grade (plus 1), whether the student was suspended during ninth grade, the GPA (based on all ninth-grade courses), and whether the student enrolled in tenth grade on time. These behaviors are strongly associated with well-known psychometric measures of noncognitive skills including the "big five" and grit.[9] Informed by Heckman et al. (2006), I use a principal component model to create a single index of these behaviors. This index is a weighted average of the non–test score outcomes and is standardized to be mean

[6] I use an algorithm to ensure high-quality matching of students to teachers. I detail this in app. A (apps. A–K are available online).

[7] Results that exclude ninth-grade repeaters entirely are essentially unchanged.

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

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

zero and unit variance. I refer to this index as the *behavior index*.[10] The behavior index has a correlation of .56 with test scores. However, analysis of variance reveals that about 75 percent of the variation in the behavior index is unrelated to test scores. As such, there is much variation in this index that is unrelated to test scores that may serve as a proxy for noncognitive skills that go largely unmeasured by standardized tests.[11]

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 2005–11 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 a private school are neither graduates nor dropouts. As such, opposite effects observed on both outcomes cannot be due to changes in private school or out-of-state enrollment. While having both measures is valuable, high school dropout is notoriously difficult to measure (Tyler and Lofstrom 2009). As such, I focus analysis on the more reliable high school graduation outcome. Roughly 4.3 percent of ninth graders are recorded as having subsequently dropped out of school, while about 82 percent graduated from high school.[12] The remaining 11 percent either transferred out of the North Carolina system or remained in school beyond the expected graduation year. Other longer-run outcome data include GPA at graduation, taking the SAT, and reported intentions to attend a 4-year college upon graduation (2006–11 cohorts). Roughly 48 percent of ninth graders took the SAT by twelfth grade, and 35 percent intended to attend a 4-year college.

To present suggestive evidence that these behaviors may proxy for skills not well measured by test scores, I examine whether these behaviors (in ninth grade) predict the longer-run outcomes conditional on test scores in ninth grade (table 2). To remove the influence of sociodemographics, all models include controls for parental education, gender, ethnicity, English and math test scores, repeater status, absences, out-of-school suspension in seventh and eighth grade, and GPA in eighth grade and include indicator variables for each secondary school. Transcript data are available only in high school so that eighth-grade GPA is observed only

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

[11] 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 (Brookhart 1993; Howley, Kusimo, and Parrott 2000) so that grades reflect a combination of skills, only some of which may be measured by test scores.

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

TABLE 1
SUMMARY STATISTICS OF STUDENT DATA

Variable	Observations	Mean	Standard Deviation	Standard Deviation within Schools	Standard Deviation within Tracks
Math z-score 8th grade[a]	573,963	.233	(.940)	(.853)	(.605)
Reading z-score 8th grade[a]	573,963	.217	(.943)	(.882)	(.678)
Repeat 8th grade	570,850	.006	(.080)	(.079)	(.073)
Suspended (8th grade)	573,963	.039	(.193)	(.191)	(.180)
Absences (8th grade)	573,963	4.593	(5.655)	(5.593)	(5.196)
GPA (8th grade)[b]	148,464	3.204	(.846)	(.515)	(.374)
Student: female	573,963	.504	(.500)	(.499)	(.479)
Student: black	573,963	.261	(.439)	(.403)	(.362)
Student: Hispanic	573,963	.072	(.259)	(.255)	(.241)
Student: white	573,963	.588	(.492)	(.446)	(.402)
Student: Asian	573,963	.021	(.142)	(.140)	(.133)
Parental education: some high school	573,963	.066	(.249)	(.246)	(.233)
Parental education: high school graduate	573,963	.394	(.489)	(.476)	(.448)
Parental education: trade school graduate	573,963	.016	(.126)	(.126)	(.122)
Parental education: community college graduate	573,963	.133	(.340)	(.338)	(.327)
Parental education: 4-year college graduate	573,963	.227	(.419)	(.410)	(.386)
Parental education: graduate school graduate	573,963	.067	(.251)	(.244)	(.230)
Number of honors classes	573,963	1.545	(1.814)	(1.602)	(.649)
Algebra I z-score (9th grade)[a]	358,315	.198	(.967)	(.898)	(.768)

English I *z*-score (9th grade)[a]	569,705	.203	(.922)	(.858)	(.645)
Geometry *z*-score (9th grade)[a]	113,693	.061	(.968)	(.832)	(.670)
Algebra II *z*-score (9th grade)[a]	34,927	.087	(.956)	(.785)	(.642)
Math *z*-score (9th grade)	477,524	.183	(.959)	(.900)	(.751)
Absences (9th grade)	573,963	3.462	(4.991)	(4.902)	(4.430)
Suspended (9th grade)	573,963	.051	(.220)	(.217)	(.202)
GPA (9th grade)	573683	2.896	(.836)	(.749)	(.581)
In 10th grade on time	573,963	.899	(.301)	(.296)	(.266)
GPA (10th grade)	421,872	2.76	(.861)	(.764)	(.620)
Dropout (2005–11 9th-grade cohorts)	531,920	.043	(.204)	(.202)	(.187)
Graduate (2005–11 9th-grade cohorts)	531,920	.824	(.381)	(.374)	(.344)
Take SAT (2006–11 9th-grade cohorts)	472,480	.477	(.499)	(.479)	(.410)
SAT total score	225,684	1,003.3	(189.0)	(165.9)	(123.7)
GPA at high school graduation	406,826	2.809	(.703)	(.646)	(.504)
Intend to attend 4-year college (2006–11 9th-grade cohorts)	472,480	.3506	(.477)	(.422)	(.349)

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

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

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

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

	Main Longer-Run Outcomes				Additional Outcomes		
	Drop Out (1)	Graduate (2)	Drop Out (3)	Graduate (4)	High School GPA at Graduation (5)	Take SAT (6)	Intend to Attend 4-Year College (7)
GPA (9th grade)	−.0353**	.0933**					
	[.000760]	[.00126]					
Log no. absences + 1 (9th grade)	.00635**	−.0198**					
	[.000317]	[.000552]					
Suspended (9th grade)	.0177**	−.0503**					
	[.00225]	[.00339]					
On time in 10th grade	−.0761**	.337**					
	[.00188]	[.00301]					
Math *z*-score (9th grade)	−.00427**	.00691**					
	[.000443]	[.000794]					
English *z*-score (9th grade)	−.00539**	.00503**					
	[.000659]	[.00112]					
Average test scores: *z*-score[a]			−.0133**	.0186**	.151**	.0465**	.0358**
			[.000747]	[.00113]	[.00151]	[.00128]	[.00125]
Behavior index: *z*-score			−.0524**	.158**	.345**	.130**	.09312**
			[.000588]	[.000781]	[.00128]	[.00073]	[.00069]
Observations	439,284	439,284	527,571	527,571	403,672	468,015	468,015

NOTE.—Robust standard errors are in brackets. In addition to school fixed effects and year fixed effects, all models include controls for student gender, ethnicity, parental education, a cubic function of math and reading test scores in seventh and eighth grade, suspension in seventh and eighth grade, days absent in seventh and eighth grade, GPA in eighth grade (for high school courses only), and whether the student had repeated seventh or eighth grade. Individuals with no eighth-grade GPA are imputed a value of 2.5, and all models include an indicator variable denoting whether the eighth-grade GPA is imputed.

[a] Where only one test-score is available, the average is the single available test score. As such, there are more observations with average scores that those with both English and math scores.

+ $p < .1$.

* $p < .05$.

** $p < .01$.

for high school courses taken while in eighth grade (about 25 percent of students).[13] Appendix F shows that the main results are robust to excluding ninth-grade GPA as a skill measure and relying on the other behaviors for which the lags are observed for all students. Columns 1 and 2 show that higher test scores in ninth grade predict less dropout and more high school graduation. Also, the non–test score behaviors in ninth grade predict variation in these outcomes conditional on test scores. The coefficients on the individual behaviors all have the expected signs and are statistically significant.

To facilitate an apples-to-apples comparison with the behavior index, I create a test score index that is the average of ninth-grade math and English scores. For both longer-run outcomes, increases in the behavior index are associated with sizable improvements conditional on test scores (cols. 3 and 4). While a 1σ increase in the test score index is associated with a 1.33 percentage point decrease in dropout, a 1σ increase in the behavior index is associated with a 5.24 percentage point decrease. Similarly, while a 1σ increase in the test score index is associated with a 1.86 percentage point increase in high school graduation, a 1σ increase in the behavior index is associated with a 15.8 percentage point increase. Columns 5–7 present patterns for high school GPA, SAT taking, and intentions to attend a 4-year college. Across all the longer-run outcomes, increases in the behavior index are associated with large and statistically significant improvements, conditional on test scores.[14] This suggests that teacher impacts on behaviors (a proxy for noncognitive skills) may be a good predictor of impacts on longer-run outcomes, above and beyond that predicted by their impacts on test scores. This is explored directly in Section VI.

III. Theoretical Framework and Model Setup

The standard value-added model assumes that student ability is one-dimensional (see Todd and Wolpin 2003). Following Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010), I extend this model so that student outcomes are functions of both cognitive and noncognitive abilities. In the model, teachers can improve skills that lead to improved longer-run outcomes but are not reflected in improved test scores. As such, teacher impacts on non–test score outcomes can provide additional information (above and beyond that contained in teacher impacts on test scores) on the extent to which they improve longer-run outcomes. For expositional purposes, I refer to students' latent competencies as *abil-*

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

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

ities, I refer to short-run student outcomes used to infer these competencies (such as test scores, course grades, etc.) as *skill measures*, and I refer to longer-run outcomes (such as high school graduation and college going) as *outcomes*.

Production of student skills.—Prior to ninth grade, each student i has a stock of cognitive and noncognitive abilities described by vector $\nu_i = (\nu_{ci}, \nu_{ni})^T$, where the subscripts c and n denote the cognitive and noncognitive dimensions, respectively.[15] This stock reflects an initial endowment and the cumulative effect of all school and parental inputs on students' incoming abilities. Each ninth-grade teacher j has a positive quality vector $\omega_j = (\omega_{cj}, \omega_{nj})^T$ describing teacher j's capacity to increase each of the two dimensions of student ability during ninth grade. Each student has a matrix given by

$$D_i = \begin{bmatrix} D_{ci} & 0 \\ 0 & D_{ni} \end{bmatrix},$$

which describes student i's responsiveness to teacher quality in each dimension. The "effective" quality of teacher j for student i (ω_{ij}) is the student matrix D_i times the underlying quality vector of teacher j given by $\omega_{ij} = D_i\omega_j$.[16]

During ninth grade, students take classes in many subjects (math, English, sciences, social studies, etc.). The two-dimensional vector φ_{i-j} represents the contribution of the ninth-grade teachers other than teacher j to the end-of-year ability of student i. Ability of student i at the end of ninth grade with teacher j is represented by the vector in (1):[17]

$$\alpha_{ij} = \upsilon_i + \omega_{ij} + \varphi_{i-j}. \tag{1}$$

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

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

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

[16] The vector ω_{ij} is a two-dimensional student-specific teacher quality vector. This relaxes the commonly made assumption that teacher effects are the same for all students (see Jackson et al. 2014).

[17] Appendix D outlines the explicit production function assumption that justifies the additive model in (1). I also present empirical evidence to support the assumption of additivity across teachers in app. J.

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

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

Teacher effects.—Teachers affect student skill measures and outcomes only through their effects on students' accumulated ability. From (2) and (3), teacher j's effect on outcome or skill measure y_z of student i, where $z \in \{s, l\}$, is a weighted average of teacher j's effective quality for each dimension of student ability $\theta_{zij} = \omega_{ij}{}^{T}\beta_z$. Let $\theta_{zj} = E[\omega_{ij}]^{T}\beta_z$ be the average effect of teacher j on outcome y_z (i.e., the effect on the average student). Because $E[\omega_{ij}] = E[D_i\omega_j]$, it follows that θ_{zj} is a linear function of the teacher quality vector (ω_j). For expositional purposes, I refer to the teachers' average effect on short-run outcomes (i.e., skill measures) as value-added.[18]

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

To illustrate this point, consider two ninth-grade skill measures, test scores (y_1) and behaviors (y_2), and a longer-run outcome, high school graduation (y_l). Assume that value-added on test scores and behaviors are perfect measures (i.e., there is no estimation or measurement error).[19] The best linear unbiased estimate of the average teacher effect on graduation (y_l) based on test score value-added (θ_{1j}) is $\gamma\theta_{1j}$, where $\gamma = \text{cov}(\theta_{lj}, \theta_{1j})/\text{var}(\theta_{1j})$. The variation in a teacher's average effect on graduation (θ_{lj}) unexplained by her test score value-added (θ_{1j}) is a linear function of her quality vector $\ddot{\theta}_{lj} = f(\omega_j)$.[20] Similarly, a teacher's behaviors value-added (θ_{2j}) unexplained by her test score value-added (θ_{1j}) is

[18] This definition is appropriate in the current context because the empirical models employed to estimate teacher effects on ninth-grade skill measures (outlined in Sec. IV) control for lagged outcomes.

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

[20] A teacher's average effect on the long-run outcome is $\theta_{lj} = \beta_{cl}\omega_{cj} + \beta_{nl}\omega_{nj}$. The variation in θ_{lj} unexplained by θ_{1j} is

$$\ddot{\theta}_{lj} = f(\omega_j) = (\beta_{cl} - \gamma\beta_{c1})\omega_{cj} + (\beta_{nl} - \gamma\beta_{n1})\omega_{nj}.$$

Similarly, the variation in θ_{2j} unexplained by θ_{1j} is

$$\ddot{\theta}_{2j} = g(\omega_j) = (\beta_{c2} - \pi\beta_{c1})\omega_{cj} + (\beta_{n2} - \pi\beta_{n1})\omega_{nj},$$

where $\pi = \text{cov}(\theta_{2j}, \theta_{1j})/\text{var}(\theta_{1j})$.

a linear function of the same teacher quality vector $\ddot{\theta}_{2j} = g(\omega_j)$. Consider the linear regression predicting the average teacher effect on the longer-run outcome (θ_{lj}) as a function of her test score value-added (θ_{1j}) and her behaviors value-added (θ_{2j}). From Greene (2002), behaviors value-added (θ_{2j}) increase the explained average teacher-level variation in graduation iff $\text{cov}(\ddot{\theta}_{lj}, \ddot{\theta}_{2j}) \neq 0$.[21] Because both $\ddot{\theta}_{lj}$ and $\ddot{\theta}_{2j}$ are functions of ω_j, it follows that $\text{cov}(\ddot{\theta}_{lj}, \ddot{\theta}_{2j}) \neq 0$, so that behaviors value-added will increase the explained teacher-level variation in graduation.[22] This argument can be applied to any additional skill measure (y_2) and any longer-run outcome (y_l). Note that this result does not require that the additional skill measure is unrelated to test scores, but only that there is meaningful variation in abilities measured by the other skill measure that is unrelated to test scores.[23]

IV. Empirical Strategy

A. *Identifying Teacher Impacts on Student Outcomes*

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

$$\begin{aligned} y_{zij} &= (v_i + \omega_{ii} + \varphi_{i-j})^T \beta_z + \varepsilon_{zij} \\ &= v_i^T \beta_z + \omega_{ij}^T \beta_z + \varphi_{i-j}^T \beta_z + \varepsilon_{zij}. \end{aligned} \tag{4}$$

[21] See app. D for a more formal proof of this statement.

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

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

Cross-multiplying out terms and substituting in θ_{zj} leads to (5):

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

When incoming student ability is not observed and one observes only the value-added of ninth-grade teacher j in a particular subject, (5) becomes

$$y_{zij} = \theta_{zj} = u_{zij}, \quad (6)$$

where

$$u_{zij} = \upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz} + \varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz} + \varepsilon_{zij}.$$

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

Because teachers and students are not distributed randomly, differences in teacher-level mean outcomes are unlikely to yield the differences in value-added of individual teachers for two reasons. First, students may sort into schools, and to teachers within schools, by parental socioeconomic status and incoming ability so that

$$E[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}|\theta_{zj}] \neq E[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}].$$

Second, good teachers may cluster in the same schools and teach the same group of students within schools because of tracking, so that

$$E[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}|\theta_{zj}] \neq E[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}].$$

For example, if good math teachers teach the same group of students as the good English teachers, average classroom outcomes for the math teacher will confound that teacher's value-added with the value-added of the English teacher to whom her students are exposed.

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

IDENTIFYING ASSUMPTION 1. Conditional random assignment of students to teachers:

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

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

IDENTIFYING ASSUMPTION 2. Conditional independence of teacher effects:

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

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

Even though

$$E\left[\hat{\theta}_{zj}\right] = \theta_{zj} + E\left[\upsilon_{ci}\beta_{cz} + \upsilon_{ni}\beta_{nz}|T_{ij}\right] + E\left[\varphi_{ci-j}\beta_{cz} + \varphi_{ni-j}\beta_{nz}|T_{ij}\right] \neq \theta_{zj},$$

under assumptions 1 and 2, $E[\hat{\theta}_{zj} - \hat{\theta}_{zj'}|T_{ij}] = \theta_{zj} - \theta_{zj'}$. That is, for a given outcome z, even with sorting of students to teachers and clustering of teachers to groups of students, in expectation, the difference in mean outcomes for teacher j and that for teacher j' conditional on T_{ij} will yield the difference in value-added between teacher j and teacher j'.

The proposed T_{ij} includes several variables. To account for student ability sorting, I include two lags of math scores, English scores, repeater status, suspensions, and attendance and a single lag of GPA.[24] To account for sorting that occurs at the group level, I include classroom averages of the eighth-grade skill measures and demographics (Protic et al. 2013). To account for sorting of teachers to groups of classes such that teacher quality may be correlated across subjects for the same student, I control for the number of honors courses taken (Aaronson, Barrow, and Sander 2007; Harris and Anderson 2012), and I include fixed effects for the student's school track (Jackson 2014). The school track is the unique combination of the 10 core academic courses, the level of math taken, and the level of English taken in a particular school.[25] Only students at the same school who also take the same academic courses, level of English, and level of math are in the same school track.[26] I refer to the school track as "track" for the remainder of the paper.

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

[25] 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 app. E for further discussion of tracks.

[26] 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 the same math or English class at different levels are in different school tracks. Because many students pursue the same course of study, less than 1 percent of all students are in singleton

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

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

B. *Identifying Teacher Impacts on Ninth-Grade Skill Measures*

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

$$y_{zicjst} = \Omega_z X_{icjst} + \tau_{st} + e_{zicjst}. \tag{9}$$

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

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

---

tracks, 83 percent of students are in tracks with more than 20 students, and the median student is in a school track with 199 other students.

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

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

The estimate $\hat{\theta}_{zj,-t}$ minimizes mean square estimation error and is an unbiased estimate of θ_{zj}. However, because $\hat{\theta}_{zj,-t}$ is estimated with noise, $\hat{\theta}_{zj,-t}$ is not the optimal out-of-sample predictor and does not minimize out-of-sample prediction errors. To minimize mean squared prediction errors, it is optimal to introduce some bias and to use the raw estimates to form empirical Bayes (or shrinkage) estimates (Gordon, Kane, and Staiger 2006; Kane and Staiger 2008; Chetty et al. 2014a).[27] This approach models the estimation error in each teacher's raw mean and shrinks noisier estimates toward the grand mean (in this case, zero). The resulting leave-year-out empirical Bayes estimate of teacher j's value-added is described by

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

This empirical Bayes estimate for each teacher's value-added is the leave-year-out teacher-level mean ($\hat{\theta}_{zj,-t}$) multiplied by λ_{zj}, an estimate of its reliability.[28] As a result, less reliable estimates (i.e., those that are estimated with more noise due to a small number of students, a small number of classrooms, or both) are shrunk toward the grand mean for all

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

[28] Following Gordon et al. (2006), Kane and Staiger (2008), Jackson and Bruegmann (2009), and Jackson (2013),

$$\lambda_{zj} = \frac{\sigma^2_{\theta_{zj}}}{\sigma^2_{\theta_{zj}} + \left\{\sum_{m_j}\left[1/\left(\sigma^2_{\varepsilon_{zcjst}} + \sigma^2_{\varepsilon_{zicjst}}/n_{cj}\right)\right]\right\}^{-1}},$$

where n_{cj} is the number of students in class c with teacher j, and m_j is the number of classrooms for teacher j. The parameters $\sigma^2_{\theta_{zj}}$, $\sigma^2_{\varepsilon_{zcjst}}$, and $\sigma^2_{\varepsilon_{zicjst}}$ are replaced by empirical estimates under the assumption

$$\text{cov}(\theta_{zj}, \varepsilon_{zcjst}) = \text{cov}(\theta_{zj}, \varepsilon_{zicjst}) = \text{cov}(\varepsilon_{zicjst}, \varepsilon_{zcjst}) = 0.$$

teachers.[29] To examine whether teacher value-added on test scores and behaviors predict teacher impacts on longer-run outcomes, I use the estimates from (11) as predictors of the longer-run outcomes. In all the empirical sections of this paper, when I refer to value-added estimates, I am referring to the leave-year-out empirical Bayes estimates as in (11).

V. Effects on Skill Measures

Before presenting impacts on longer-run outcomes, I examine the magnitudes of teacher value-added on the proposed skill measures (i.e., test scores and behaviors). I follow Kane and Staiger (2008) and for each outcome use the covariance between mean classroom residuals for the same teacher as a measure of the variance of the persistent component of teacher value-added ($\hat{\sigma}^2_{\theta_{zj}}$).[30] The square roots of estimated variances (i.e., the implied standard deviations of teacher value-added) for all ninth-grade outcomes are presented for each subject in table 3.

The standard deviation of the math teacher value-added on math test scores is 0.084σ so that having a math teacher with value-added at the 85th versus 50th percentile on math test scores would increase math scores by roughly 0.084σ. The relationship between average test scores and graduation in column 4 of table 2 implies that this would be associated with a 0.16 percentage point increase in the likelihood of high school graduation. Looking to value-added on behaviors, having a math teacher at the

---

Under this assumption, $\text{var}(e_{zicjst}) = \sigma^2_{\bar{e}_{sujst}} + \sigma^2_{\bar{e}_{scjst}} + \sigma^2_{\theta zj}$ and $\text{cov}(\bar{e}_{zcjt}, \bar{e}_{zc'jt'}) = \sigma^2_{\theta zj}$, where $\bar{e}_{zcjt}$ is the average residual for classroom c for teacher j in year t and $\bar{e}_{zc'jt'}$ is the average residual for classroom c' for teacher j not in year t. As such, $\sigma^2_{\bar{e}_{sujst}}$, the empirical estimate of the variance of the student-level errors, is estimated using the sample variance of the student-level residuals within classrooms. Also $\sigma^2_{\theta_{zj}}$, the empirical estimate of the variance of the true teacher value-added on outcome z, is estimated using the sample covariance of classroom-level mean residuals for the same teacher in different years. Under the assumptions above, I can obtain an empirical estimate of $\sigma^2_{\bar{e}_{scjst}}$, the variance of the classroom-level shocks, using the variance of the total residual, $\text{var}(e_{zicjst})$, minus the empirical estimates of $\sigma^2_{\bar{e}_{sujst}}$ and $\sigma^2_{\theta_{zj}}$.

[29] Teachers with no estimated raw fixed effects (i.e., those in the data for only 1 year) are shrunk toward the mean of other teachers with similar observable attributes. Teachers with missing estimates are given the fitted value from a regression predicting $\hat{\mu}_{zjt}$ based on observable teacher characteristics (gender, ethnicity, experience, certification, license status, college selectivity, and test scores). Teachers for whom there are no observable characteristics are given the mean of the distribution of the estimated $\hat{\mu}_{zjt}$. Results are very similar to those obtained when the teacher estimates are shrunk to zero for teachers with no estimated out-of-sample effect.

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

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

	ALL TEACHERS							
	English Score	Math Score	Suspended	Absences[a]	9th-Grade GPA	In 10th Grade on Time	Behaviors Index	10th-Grade GPA
SD:								
English teachers	.0301	.0292	.0104	.0434	.0415	.0212	.0552	.0360
Math teachers	.0204	.0844	.0121	.0001	.0632	.0264	.0801	.0501
All teachers	.018	.0751	.0108	.02839	.0446	.0247	.0769	.0315

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

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

85th versus 50th percentile reduces the likelihood of being suspended by 1.2 percentage points, has no impact on absences, increases GPA by 0.063 grade points, and increases on-time grade progression by 2.64 percentage points. Combining the ninth-grade behaviors into a single variable, having a math teacher at the 85th versus 50th percentile of value-added on the behavior index would increase the behavior index by 0.08σ. The relationships in table 2 suggest that this would lead to a 1.27 percentage point increase in the likelihood of high school graduation. Patterns for English teachers are similar. However, as in other settings, value-added on English scores are smaller than those on math scores (see Jackson et al. 2014). The correlations indicate that having an English teacher with value-added at the 85th versus 50th percentile on English scores would increase English scores by 0.03σ. Having an English teacher with value-added at the 85th versus 50th percentile on the behavior index would increase the behavior index by roughly 0.055σ—an effect size on behaviors that is on the same order of magnitude as those for math teachers. The patterns presented in table 3 indicate that there is economically meaningful variation in outcomes across teachers that persists across classrooms.

One may worry that these correlations are driven by systematic reporting bias (e.g., teachers who are easy graders or do not report students to the principal's office may mechanically appear to improve student outcomes without actually improving underlying behaviors). Because passing English and math is required to graduate from high school and an

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

Relationship between teacher effects across skill measures.—To explore whether teachers who improve test scores improve other skill measures, table 4 presents the raw correlations between the value-added estimates on the different skill measures, where the data for both math and English teachers are combined. Teachers with higher test score value-added are associated with better non–test score outcomes, but the relationships are weak. The correlations between test score value-added and that on being suspended or absences are both below .1. The test score value-added estimates are somewhat more highly correlated with value-added on GPA ($r = .22$) and on-time grade progression ($r = .16$), but not strongly so. The correlation between test score value-added and that on the behavior index is only .15 such that less than 3 percent of the variation in teacher value-added on the behavior index is associated with teacher value-added on test scores, and vice versa. Looking to impacts on tenth-grade GPA (which is free from reporting and grading biases), the correlation with test score value-added is only .11. However, because the value-added estimates are estimated with noise, the variation in value-added on behaviors that is unrelated to value-added on test scores may simply reflect statistical noise, and not systematic variation in teacher quality per se.

To further explore whether teachers' behaviors value-added reflect impacts on skills that are unmeasured by their test score value-added,

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

	TEACHER VALUE-ADDED						
	Test Score	Suspended	Absences[a]	GPA	In 10th Grade on Time	Behavior Index	10th-Grade GPA
Teacher value-added:							
Test score	1						
Suspended	−.0726	1					
Absences[a]	−.0390	.0983	1				
GPA	.2266	−.1454	−.0863	1			
In 10th grade on time	.1610	−.1185	−.0642	.3822	1		
Behavior index	.1494	−.3325	−.4461	.5716	.5454	1	
10th-grade GPA	.1147	−.0449	−.0601	.3471	.0973	.2320	1

NOTE.—This table reports the estimated two-way correlation coefficient between the estimated teacher value-added ($\hat{\mu}_{zjt}$) on each skill measure and each other skill measure.
[a] Absences refer to the natural log of the number of absences plus one.

I regress student skill measures on their teachers' leave-year-out value-added estimates for those skill measures. Specifically, I estimate equation (12), where all variables are defined as in (9) and $\hat{\mu}_{\text{test},jt}$ and $\hat{\mu}_{\text{behavior},jt}$ are the leave-year-out empirical Bayes value-added estimates on test scores and the behaviors, respectively:

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

For ease of interpretation, the teacher value-added estimates are multiplied by scaling factors ϱ_1 and ϱ_2 so that the coefficients δ_1 and δ_2 identify the effect of increasing teacher value-added on test scores and the behaviors, respectively, by one standard deviation (as presented in table 3).[31] Data for all subjects are stacked, and the results are presented for both subjects combined. All models include indicators for the specific course of the teacher (I_d) (i.e., English, geometry, algebra I, and algebra II). To

[31] To obtain the scaling index for each outcome I first estimate equation (a) below for each outcome z:

$$y_{zicjst} = \beta_z X_{icjst} + \pi_z \cdot \hat{\mu}_{zjt} + v_{zicjst}. \tag{a}$$

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

account for the fact that individual students enter the stacked data set in both subjects and individual teachers have multiple students, standard errors are adjusted for two-way clustering at the teacher and student levels following Cameron, Gelbach, and Miller (2011).

Table 5 presents the coefficients on the rescaled value-added estimates. As expected, columns 1, 7, and 10 show that teachers who raise a given skill measure out of sample have large statistically significant effects on that same skill measure. Increasing teacher test score value-added (across both subjects) by one standard deviation increases test scores by 0.0685σ (p-value $< .01$),[32] increasing teacher behaviors value-added by one standard deviation increases the behavior index by 0.0579σ (p-value $< .01$), and increasing teacher tenth-grade GPA value-added by one standard deviation increases tenth-grade GPA by 0.0357σ (p-value $< .05$). Consistent with the teacher value-added on the skill measures being positively correlated, columns 2, 6, and 9 reveal that teachers who raise test scores improve behaviors and vice versa.

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

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

This section tests whether teachers who improve behaviors cause improved longer-run outcomes (conditional on their test score value-added). To this aim, I estimate equation (12) in which the outcomes are measures of high school completion, whether the students subsequently dropped out of secondary school by twelfth grade, and whether they graduated from high school. For ease of interpretation, I present estimates of the average

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

TABLE 5
EFFECTS OF TEACHER VALUE-ADDED ON SHORT-RUN SKILL MEASURES

	STUDENT TEST SCORE IN 9TH GRADE					STUDENT BEHAVIORS IN 9TH GRADE			STUDENT GPA IN 10TH GRADE		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Teacher value-added:											
9th-grade test score	.0685**		.0690**		.0685**	.0074**		.0061**	.0042**		.0038**
	[.0028]		[.0028]		[.00281]	[.0016]		[.0016]	[.0012]		[.0013]
9th-grade behaviors		.0274*	−.0214+				.0579**	.0536**			
		[.0124]	[.0128]				[.0106]	[.0107]			
10th-grade GPA				.0987**	.0012					.0357*	.0298*
				[.0230]	[.0206]					[.0147]	[.0149]
School track effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	942,291	942,291	942,291	942,291	942,291	942,291	942,291	942,291	728,529	728,529	728,529

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

+ $p < .1$.
* $p < .05$
** $p < .01$

marginal effects using linear probability models. Because linear models can be misleading about marginal effects for binary outcomes, I also present conditional logit estimates and the ensuing implied marginal effects. To quantify the increase in the ability to predict variability in teachers' longer-run impacts by adding behaviors value-added, using the linear model, I estimate (12) both with and without behaviors value-added, and I compute the percentage increase in the predicted variance of the teacher effects on the longer-run outcomes.[33] The results are presented in table 6.

Column 1 presents the effect of increasing test score value-added on high school graduation when behaviors value-added is not included. On average, one standard deviation higher test score value-added leads to a 0.152 percentage point increase in high school graduation (p-value < .01). To put this estimated effect into perspective, the linear relationship between a one standard deviation increase in test scores and graduation in table 2 (1.86 percentage points) multiplied by the estimated standard deviation of test score value-added in table 3 (0.075) implies that a one standard deviation increase in teacher test score value-added would increase high school graduation by $1.86 \times 0.075 = 0.139$ percentage points. This is very close to the estimated magnitudes—suggesting that the results are reasonable.

Column 2 presents the teacher effects on high school graduation using teacher value-added on both the behavior index and test scores. Given that the two are weakly correlated, the point estimate for test score value-added remains largely unchanged. Conditional on a teacher's behaviors value-added, increasing test score value-added by one standard deviation increases high school graduation by 0.118 percentage points, and conditional on a teacher's test score value-added, increasing a teacher's behaviors value-added by one standard deviation increases high school graduation by 1.46 percentage points (p-value < .01). The linear relationship between a one standard deviation increase in behaviors and graduation in table 2 (15.8 percentage points) multiplied by the estimated standard deviation

[33] Specifically, I estimate the following two equations:

$$y_{zicjst} = \Omega_z X_{it} + \delta_{z1} \cdot \left(\varrho_1 \hat{\mu}_{\text{test},jt}\right) + \delta_d \sum_{d=1}^{4} I_d + v_{zicjst}, \tag{b}$$

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

I compute the variance of the fitted values for each teacher from both models. In models without behaviors the value-added, i.e., (b), this is $a = \text{var}(\hat{\delta}_1 \cdot (\varrho_1 \hat{\mu}_{test,jt}))$, and in models with teacher value-added on both, i.e., (c), this is

$$b = \text{var}\left[\hat{\delta}_1 \cdot \left(\varrho_1 \hat{\mu}_{\text{test},jt}\right) + \hat{\delta}_2 \cdot \left(\varrho_2 \hat{\mu}_{\text{behavior},jt}\right)\right].$$

The percentage increase in explained variance from also including behaviors value-added (vs. using their test score value-added alone) is $100 \times [(b \div a) - 1]$.

TABLE 6
EFFECTS OF TEACHER VALUE-ADDED ON HIGH SCHOOL COMPLETION

	Student: Graduate from High School					Student: Drop Out of High School				
	Linear Probability Model			Conditional Logit[a]		Linear Probability Model			Conditional Logit[a,b]	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Teacher value-added:										
9th-grade test score	.0015**	.0012*	.0013*	.0088	.01	−.0004	−.0003	−.0004	−.0055	−.0084
	[.0005]	[.00054]	[.0005]	[.0064]	[.0064]	[.0003]	[.0003]	[.0003]	[.0099]	[.0101]
				(.002)	(.0023)				(−.0012)	(−.0018)
9th-grade behaviors		.0146**		.1442**			−.0041*		−.128*	
		[.00319]		[.0343]			[.0019]		[.0583]	
				(.0331)					(−.0271)	
10th-grade GPA			.0146**		.162**			−.0031		−.0618
			[.0056]		[.0637]			[.0031]		[.0996]
					(.0375)					(−.012)
% increase in explained variance		305%	97%				326%	59%		

School track effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	891,868	891,868	891,868	579,512	579,512	891,868	891,868	891,868	570,390	570,390

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

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

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

$^{+}$ $p < .1$.

* $p < .05$.

** $p < .01$.

of teacher behaviors value-added in table 3 (0.0769) implies that a one standard deviation increase in a teacher's behaviors value-added would increase students' high school graduation by $15.8 \times 0.0769 = 1.215$ percentage points. This is very close to the estimated magnitudes, suggesting that the results are reasonable and that the magnitudes are plausible. Comparing the estimated teacher-level variability in high school graduation from the fitted models with both value-added estimates to those using only test score value-added, including behaviors value-added increases the explained variance of teacher effects on graduation by 305 percent, that is, more than quadruples the variance of the identifiable teacher effect on high school graduation. This is consistent with results by Chamberlain (2013), who finds that test score value-added may account for less than one-fifth of the overall effect of teachers on college going.

Column 4 presents the results from a conditional logit specification to more accurately reflect the binary outcome. The estimated coefficient estimates are presented, with standard errors below in brackets, and the average marginal effects are presented in parentheses below that.[34] In models with teacher value-added on both test scores and behaviors, increasing test score value-added by one standard deviation increases high school graduation by 0.2 percentage points (p-value > .1), and increasing the teacher's behaviors value-added by one standard deviation increases high school graduation by 3.31 percentage points (p-value < .01). Even though the linear and nonlinear models yield somewhat different marginal effects, they are on the same order of magnitude and have overlapping 95 percent confidence intervals. To put these effect sizes into perspective, consider the following back-of-the-envelope calculation. In the linear model, increasing a teacher's behaviors value-added by one standard deviation increases high school graduation by 1.46 percentage points, on average. The average teacher has 54.5 students a year. According to the Bureau of Labor Statistics (US Department of Labor 2016), completing high school is associated with $11,000 higher annual earnings. Assuming this difference is causal, increasing high school graduation rates by 1.46 percentage points would increase annual earnings by roughly $160 per year per student. This figure multiplied by 54 students is $8,670 higher cohort earnings each year. Assuming this increase stays the same each year for 40 years, at a 7 percent discount rate, this translates into $126,286 in present discounted lifetime earnings per year of students taught. Even though some of the raw differences in earnings assumed in this rough calculation may reflect selection, under most reasonable assumptions regarding the economic

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

benefits of completing high school, the estimated effects are economically important.

The other measure of school completion is high school dropout. High school dropout is notoriously difficult to measure (Tyler and Lofstrom 2009) so that the effects will likely be muted. However, it is helpful to show that the same patterns hold for both high school graduation and high school dropout. Column 7 shows that, on average, a one standard deviation increase in teacher test score value-added reduces the likelihood of dropping out by 0.03 percentage points, and a one standard deviation increase in teacher behaviors value-added reduces the likelihood of dropout by 0.4 percentage points (p-value < .05). While the point estimates from the linear model are smaller than those for graduation, the implied marginal effects from the conditional logit model are similar across the two outcomes. In models with value-added on both test scores and behaviors (col. 9), a one standard deviation increase in teacher value-added on test scores and behaviors reduces the likelihood of dropout by 0.12 and 2.71 percentage points, respectively. As with high school graduation, including teachers' behaviors value-added increases the explained teacher-level variance in dropout by 326 percent. The consistency across both measures of school completion suggests that the estimated effects reflect real changes in human capital acquisition. Note that if teachers affect skills not captured by test scores or the behaviors (which is likely), the estimates presented may still understate teachers' full effect on longer-run outcomes.[35]

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

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

real improvement in student skills or promoting productive behaviors that improve students' longer-run outcomes.[36]

A. *Testing the Identifying Assumptions*

The first identifying assumption is that students are randomly assigned to teachers conditional on observables.[37] To present evidence that this condition is satisfied, I first implement a test for selection on observables (app. G). I show that conditional on eighth-grade outcomes and controls for tracks, teacher value-added estimates are unrelated to their students' predicted dropout and predicted graduation (weighted indices of parental education, gender, and ethnicity and seventh-grade math scores, reading scores, grade repetition, suspensions, and absences). To test for selection on unobservables within school cohorts, I follow Chetty et al. (2014b) and exploit the statistical fact that the effects of any selection among students within a cohort at a given school will be eliminated by aggregating the treatment to the school year level and relying only on cohort-level variation across years within schools. That is, if value-added estimates merely capture selection within school cohorts, then the arrival of a teacher who increases the average teacher estimated value-added for a cohort but has no effect on real teacher quality should have no effect on average student outcomes for that cohort. Conversely, if the value-added estimates reflect real impacts, differences in average estimated teacher value-added across cohorts (driven by changes in teaching personnel within schools over time) should be associated with similar outcome differences as similar differences in estimated value-added across individual students within cohorts. To test for this, I implement instrumental variables models that use only variation across cohorts within a school (app. G). The main findings are robust to using the clean variation across cohorts. In sum, I find no evidence of selection bias so that the first identifying assumption is likely valid.

The second identifying assumption is that, conditional on observables, the quality of a student's teacher in one subject is unrelated to the quality

[36] Appendix F presents results using teacher value-added on each behavior individually. Teacher value-added on individual behaviors have the expected signs, and many are statistically significant. Because eighth-grade GPA is imperfectly measured, I show that the results are robust to using teacher value-added on an index that excludes GPA as a skill measure entirely. In sum, app. F shows that the value-added on no single behavior drives the effects and that it is the shared variability across the behaviors (which I posit is due to noncognitive skills).

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

of that student's teachers in other subjects. I test this assumption in two ways (app. G). First, for each student I correlate the estimated math and English teacher value-added. The correlation between the math and English teacher test score value-added is .008, that for math and English teacher behaviors value-added is .0078, and that for math and English teacher effects tenth-grade GPA value-added is .0087. In a regression predicting the math teacher's value-added as a function of the English teacher's value-added, all coefficients are close to zero and all have p-values larger than .1. I also test whether the main results are robust to the inclusion of fixed effects for the other subject teachers and school-by-year fixed effects (to account for subject teachers other than math and English).[38] Linear probability models that include other subject teacher fixed effects and school-by-year fixed effects are almost identical to those that do not. This is consistent with no conditional correlation between the quality of teachers across subjects, suggests that the empirical strategy isolates the contribution of the individual teachers, and suggests that the second identifying assumption is valid.

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

While high school dropout and graduation are the main longer-run outcomes in this study, I present effects of ninth-grade teachers on a few intermediate outcomes and measures of college going (table 7). I focus attention on the impacts of teachers' behavior value-added conditional on test score value-added. Consistent with the graduation and dropout results, conditional on teachers' test score value-added, a one standard deviation increase in behaviors value-added increases enrolling in tenth grade by 2 percentage points (p-value < .1), increases tenth-grade GPA by 0.013 grade points (p-value < .1), increases SAT taking by 1.16 percentage points (p-value < .1), increases the likelihood of reporting plans to attend a 4-year college after high school graduation by 1.03 percentage points (p-value < .05), and increases graduating high school GPA by 0.0214 points (p-value < .01). The one outcome for which behaviors value-added adds no explanatory power is total SAT score (which is affected by a teacher's test score value-added).

Similarly to the patterns for high school completion, including teachers' behaviors value-added increases the identifiable teacher-level variance by 793 percent for tenth-grade enrollment, 33 percent for tenth-grade GPA, 305 percent for graduation, 326 percent for dropout, 228 percent for SAT taking, 193 percent for 4-year college intentions, and 607 percent for high

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

TABLE 7
EFFECTS OF TEACHER VALUE-ADDED ON VARIOUS LONG-TERM OUTCOMES

	Enrolled in 10th Grade (1)	10th-Grade GPA (2)	Drop Out of School (3)	Graduate from High School (4)	Take the SAT (5)	Total SAT (6)	Intend to Attend 4-Year College[a] (7)	GPA in 12th Grade (8)
Teacher value-added:								
9th-grade test score	.000836^{+}	.00389**	−.000315	.00118*	.00114^{+}	.596*	.00115	.00109
	[.000498]	[.00120]	[.000290]	[.000546]	[.000686]	[.274]	[.000801]	[.000873]
9th-grade behaviors	.0204**	.0130^{+}	−.00407*	.0146**	.0116**	−.232	.01031*	.0214**
	[.00318]	[.00786]	[.00192]	[.00319]	[.00378]	[1.765]	[.00442]	[.00566]
% increase in explained variance	793%	33%	326%	305%	228%	.10%	193%	607%
Teacher value-added:								
9th-grade test score	.00116*	.00375**	−.000363	.00130*	.00129^{+}	.596*	.00128	.00129
	[.000521]	[.00119]	[.000292]	[.000556]	[.000696]	[.275]	[.000811]	[.000877]
10th-grade GPA	.00974*	.0298*	−.00402	.0146**	.00796	−.319	.01142	.0190*
	[.00495]	[.0149]	[.00305]	[.00565]	[.00701]	[3.008]	[.00865]	[.00905]
% increase in explained variance	57%	54%	59%	97%	35%	.25%	88%	151%
Observations	942,291	728,529	891,868	891,868	789,627	401,744	789,627	701,813

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

[a] Note that intentions to attend college are available only for the 2006–11 cohorts.

$^{+}$ $p < .1$.

* $p < .05$.

** $p < .01$.

school GPA. The lower panel presents results using value-added on tenth-grade GPA to assuage concerns regarding reporting biases and mechanical effects. The results are less precise but similar to those found using ninth-grade behaviors value-added. In sum, teachers' behaviors value-added improve the ability to identify teachers who improve a variety of longer-run outcomes considerably.[39]

C. *Possible Policy Uses of Effects on Behaviors*

I briefly discuss potential applications of teacher behaviors value-added to policy making. One possibility would be to identify observable teacher characteristics associated with behaviors value-added and select teachers with these characteristics. To determine the scope of this type of policy, I regress the behavior index on observable teacher characteristics while controlling for school tracks, year effects, and student covariates (app., table I1). While observable teacher characteristics predict effects on test scores, none of the observable teacher characteristics—years of teaching experience, full certification, teaching exams scores, regular licensing, and college selectivity (as measured by the 75th percentile of the SAT scores at the teacher's college)—are significantly related to behaviors.[40] However, this does not preclude the use of more detailed teacher information to better predict teacher effects on a broad range of skills.

Another policy application is to provide incentives for teachers to improve behaviors. However, because some of the behaviors can be "improved" by changes in teacher behavior that do not improve student skills or behaviors (such as inflating grades and misreporting misconduct), attaching external stakes to the behavior index may not improve student skills. There are three feasible solutions to this "gameability" problem. One possibility is to find measures of noncognitive skills that are difficult to adjust unethically. For example, classroom observations and student and parent surveys may provide valuable information about student skills not measured by test scores and are less easily manipulated by teachers. One could attach external incentives to both these measures of noncognitive skills and test scores to promote better longer-run outcomes. Another approach is to provide teachers with incentives to improve the behaviors of students in their classrooms the following year, when the teacher's influence may

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

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

still be present, but the teacher can no longer manipulate student behaviors (as in Carrell and West [2010] and Figlio, Schapiro, and Soter [2015]). A final solution is to identify teaching practices that improve behaviors and provide incentives for teachers to engage in these practices. Such approaches have been used successfully to increase test scores (Allen et al. 2011; Taylor and Tyler 2012). In sum, the teacher effects on the behaviors 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 skills. In this model, teachers may have effects on skills that affect longer-run outcomes, are not reflected in test scores, but are reflected in other skill measures. I use an index of behaviors to proxy for noncognitive skills and find that ninth-grade teachers have meaningful impacts (i.e., value-added) on both test scores and these behaviors. While test scores and behaviors are positively correlated, value-added on behaviors explain significant variability in teacher impacts on high school graduation and dropout that are not captured by their test score value-added. Adding teachers' behaviors value-added more than doubles the identifiable teacher-level variability on longer-run outcomes such as high school graduation, SAT taking, and intentions to attend college.

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

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

References

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

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

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

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

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

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

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

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

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

Cameron, Colin, Jonah Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multi-way Clustering." *J. Bus. and Econ. Statis.* 21 (2): 238–49.

Carneiro, P., C. Crawford, and A. Goodman. 2007. "The Impact of Early Cognitive and Non-cognitive Skills on Later Outcomes." Discussion Paper no. 0092, Centre Econ. Educ., London School Econ. and Polit. Sci.

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *J.P.E.* 118 (3): 409–32.

Chamberlain, Gary. 2013. "Predictive Effects of Teachers and Schools on Test Scores, College Attendance, and Earnings." *Proc. Nat. Acad. Sci.* 110 (October 22): 17176–82.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *A.E.R.* 104 (9): 2593–2632.

———. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *A.E.R.* 104 (9): 2633–79.

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

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

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

———. 2011. "Better Schools, Less Crime?" *Q.J.E.* 126 (4): 2063–2115.

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

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

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *J. Personality and Soc. Psychology* 92 (6): 1087–1101.

Ehrenberg, R. G., D. D. Goldhaber, and D. J. Brewer. 1995. "Do Teachers' Race, Gender, and Ethnicity Matter? Evidence from NELS88." *Indus. and Labor Relations Rev.* 48:547–61.

Figlio, David N., Morton O. Schapiro, and Kevin B. Soter. 2015. "Are Tenure Track Professors Better Teachers?" *Rev. Econ. and Statis.* 97 (4): 715–24.

Fredriksson, P., B. Ockert, and H. Oosterbeek. 2013. "Long-Term Effects of Class Size." *Q.J.E.* 128 (1): 249–85.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Discussion Paper no. 2006-01, Hamilton Project, Brookings Inst., Washington, DC.

Greene, William. 2002. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.

Harris, D., and A. Anderson. 2012. "Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence from Middle School Teachers." Panel paper, Assoc. Public Policy Analysis and Management, Washington, DC.

Heckman, James J. 1981a. "Heterogeneity and State Dependence." In *Studies in Labor Markets*, edited by Sherwin Rosen, 91–140. Chicago: Univ. Chicago Press (for NBER).

———. 1981b. "Statistical Models for Discrete Panel Data." In *Structural Analysis of Discrete Data and Econometric Applications*, edited by Charles F. Manski and Daniel L. McFadden. Cambridge, MA: MIT Press.

———. 1999. "Policies to Foster Human Capital." Working Paper no. 7288, NBER, Cambridge, MA.

Heckman, James J., John Eric Humphries, and Gregory Veramendi. 2016. "Dynamic Treatment Effects." *J. Econometrics* 191 (2): 276–92.

Heckman, James J., and T. Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Econ.* 19 (4): 451–64.

Heckman, James J., Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *A.E.R.* 103 (6): 2052–86.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *A.E.R.* 91 (2): 145–49.

Heckman, James J., J. Stixrud, and S. Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *J. Labor Econ.* 24 (3): 411–82.

Howley, A., P. S. Kusimo, and L. Parrott. 2000. "Grading and the Ethos of Effort." *Learning Environments Res.* 3:229–46.

Jackson, C. Kirabo. 2013. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers." *Rev. Econ. and Statis.* 95:1096–1116.

———. 2014. "Teacher Quality at the High School Level: The Importance of Accounting for Tracks." *J. Labor Econ.* 32 (4): 645–84.

Jackson, C. Kirabo, and E. Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Econ. J.: Appl. Econ.* 1 (4): 85–108.

Jackson, Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher Related Policies." *Ann. Rev. Econ.* 6:801–25.

Jacob, Brian, Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *J. Human Resources* 45 (4): 915–43.

Jennings, J. L., and T. A. DiPrete. 2010. "Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology Educ.* 83 (2): 135–59.
John, O., A. Caspi, R. Robins, T. Moffit, and M. Stouthamer-Loeber. 1994. "The 'Little Five': Exploring the Nomological Network of the Five-Index Model of Personality in Adolescent Boys." *Child Development* 65:160–78.
Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Measures of Effective Teaching Project Research Paper, Bill and Melinda Gates Found., Seattle.
Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper no. 14607, NBER, Cambridge, MA.
Kautz, Tim, and Wladimir Zanoni. 2014. "Measuring and Fostering Non-cognitive Skills in Adolescence: Evidence from Chicago Public Schools and the OneGoal Program." Manuscript, Univ. Chicago.
Koedel, C. 2008. "Teacher Quality and Dropout Outcomes in a Large, Urban School District." *J. Urban Econ.* 64 (3): 560–72.
Lindqvist, E., and R. Vestman. 2011. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Econ. J.: Appl. Econ.* 3 (1): 101–28.
Lleras, Christy. 2008. "Do Skills and Behaviors in High School Matter? The Contribution of Noncognitive Factors in Explaining Differences in Educational Attainment and Earnings." *Soc. Sci. Res.* 37:888–902.
Lounsbury, J. W., R. P. Steel, J. M. Loveland, and L. W. Gibson. 2004. "An Investigation of Personality Traits in Relation to Adolescent School Absenteeism." *J. Youth and Adolescence* 33 (5): 457–66.
Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood. 2013. "A Composite Estimator of Effective Teaching." Measures of Effective Teaching Project Research Paper, Bill and Melinda Gates Found., Seattle.
Protik, Ali, Elias Walsh, Alexandra Resch, Eric Isenberg, and Emma Kopa. 2013. "Does Tracking of Students Bias Value-Added Estimates for Teachers?" Working paper, Mathematica Policy Res., Princeton, NJ.
Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.
Rothstein, J. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Q.J.E.* 125 (1): 175–214.
Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *A.E.R.* 102 (7): 3628–51.
Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Econ. J.* 113:F3–F33.
Tyler, John H., and Magnus Lofstrom. 2009. "Finishing High School: Alternative Pathways and Dropout Recovery." *Future of Children* 19 (1): 77–103.
US Department of Labor, Bureau of Labor Statistics. 2016. "Employment Projections: Earnings and Unemployment Rates by Educational Attainment, 2015." Bur. Labor Statis., Washington, DC. http://www.bls.gov/emp/ep_chart_001.htm.
Waddell, G. 2006. "Labor-Market Consequences of Poor Attitude and Low Self-Esteem in Youth." *Econ. Inquiry* 44 (1): 69–97.