NBER WORKING PAPER SERIES

NON-COGNITIVE ABILITY, TEST SCORES, AND TEACHER QUALITY: EVIDENCE FROM 9TH GRADE TEACHERS IN NORTH CAROLINA

C. Kirabo Jackson

Working Paper 18624
http://www.nber.org/papers/w18624

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2012

I thank David Figlio, Jon Guryan, Simone Ispa-Landa, Clement Jackson, Mike Lovenheim, James Pustejovsky, Jonah Rockoff, Dave Deming, and Steven Rivkin for insightful comments. I also thank Kara Bonneau from the NCERDC and Shayna Silverstein. This research was supported by funding from the Smith Richardson Foundation. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina

C. Kirabo Jackson

NBER Working Paper No. 18624
December 2012, Revised October 2014
JEL No. H0,I2,J0

ABSTRACT

This paper presents a model where teacher effects on long-run outcomes reflect effects on both cognitive skills (measured by test-scores) and non-cognitive skills (measured by non-test-score outcomes). Consistent with the model, results from administrative data show that teachers have causal effects on skills not measured by testing, but reflected in absences, suspensions, grades, and on-time grade progression. Teacher effects on these non-test-score outcomes in 9th grade predict longer-run effects on high-school completion and proxies for college-going—above and beyond their effects on test scores. Effects on non-test-score outcomes are particularly important for English teachers for whom including effects on the non-test-score outcomes triples the predicable variability of teacher effects on longer-run outcomes.

C. Kirabo Jackson
Northwestern University
School of Education and Social Policy
2040 Sheridan Road
Evanston, IL 60208
and NBER
kirabo-jackson@northwestern.edu

Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina[1]

C. Kirabo Jackson, 30 July, 2014
Northwestern University and NBER

This paper presents a model where teacher effects on long-run outcomes reflect effects on both cognitive skills (measured by test-scores) and non-cognitive skills (measured by non-test-score outcomes). Consistent with the model, results from administrative data show that teachers have causal effects on skills not measured by testing, but reflected in absences, suspensions, grades, and on-time grade progression. Teacher effects on these non-test-score outcomes in 9th grade predict longer-run effects on high-school completion and proxies for college-going—above and beyond their effects on test scores. Effects on non-test-score outcomes are particularly important for English teachers for whom including effects on the non-test-score outcomes triples the predicable variability of teacher effects on longer-run outcomes. (JEL I21, J00)

> *"The preoccupation with cognition and academic "smarts" as measured by test scores to the exclusion of social adaptability and motivation causes a serious bias in the evaluation of many human capital interventions"* (Heckman, 1999).

There is a general consensus that non-cognitive skills not captured by standardized tests, such as adaptability, self-restraint, and motivation, are important determinants of adult outcomes (Lindqvist & Vestman, 2011; Heckman & Rubinstein, 2001; Borghans, Weel, & Weinberg, 2008; Waddell, 2006). Also, interventions that have no effect on test scores have meaningful effects on long-term outcomes, such as educational attainment, earnings, and crime (Booker et al. 2011; Deming, 2009; Deming, 2011).[2] This suggests that schooling produces both cognitive skills (measured by standardized tests) and non-cognitive skills (reflected in socio-behavioral outcomes), both of which determine adult outcomes. Accordingly, evaluating interventions based on test scores may capture only one dimension of the skills required for adult success, and more comprehensive evaluations of interventions "*would account for their effects on producing the noncognitive traits that are also valued in the market*" (Heckman & Rubinstein, 2001).

Policy makers, educators, parents, and researchers agree that teachers are an important component of the schooling environment. Studies show that having a teacher at the 85th percentile

[1] I thank David Figlio, Jon Guryan, Simone Ispa-Landa, Clement Jackson, Mike Lovenheim, James Pustejovsky, Jonah Rockoff, Dave Deming, and Steven Rivkin for insightful comments. I also thank Kara Bonneau from the NCERDC and Shayna Silverstein. This research was supported by funding from the Smith Richardson Foundation.
[2] Heckman, Pinto, & Savelyev (forthcoming) also find that changes in personality traits explain the positive effect of the Perry Preschool Program on adult outcomes.

of the quality distribution (as measured by student test scores) versus the 15th percentile is associated with between 8 and 20 percentile points higher scores in math and reading (Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005). While economists do not care about test scores *per se*, the focus on test scores occurs because they are often the best available measure of student skills. However, the research on non-cognitive skills provides reason to suspect that teacher effects on test scores may fail to capture teachers' overall effects. Several districts publicly release estimates of teachers' average effects on student test scores (value-added) and use them in hiring and firing decisions. Accordingly, it is important that these measures reflect teachers' effects on long-run outcomes, not *only* their effect on cognitive ability.

To speak to this issue, this research tests whether teachers have causal effects on both test scores and a proxy for non-cognitive ability (a weighted average of absences, suspensions, course grades, and on-time grade progression). It also investigates whether teachers who improve test scores also improve these non-test score outcomes. Finally, it tests whether 9th grade teacher effects on this proxy for non-cognitive skills predict effects on longer-run outcomes (e.g. high school completion and intentions to attend college) conditional on test score effects. It then assesses the extent to which test score measures understate the overall importance of teachers. This paper presents the first analysis of teacher effects on both cognitive and non-cognitive outcomes, and is the first to investigate whether teacher effects on non-cognitive outcomes predict teacher effects on important longer-run effects that would go undetected by test score value-added alone.[3]

Opponents of using test scores to infer teacher quality have raised two concerns. The first is that improvements in test scores do not necessarily indicate better long-run outcomes; teachers might engage in grade-inflating practices and those skills measured by test scores may not be associated with improved long-term outcomes. Chetty, Friedman, & Rockoff (2011) assuage this concern by demonstrating that teachers who improve test scores also improve students' outcomes into adulthood. The second concern is that student ability is multidimensional, while test scores

[3] In existing work, Alexander, Entwisle, & Thompson (1987), Ehrenberg, Goldhaber, & Brewer (1995) and Downey & Shana (2004) find that students receive better teacher evaluations of behavior when students and teachers are more demographically similar, and Jennings & DiPrete (2010) finds that certain kindergarten classrooms are associated with meaningful differences in teacher evaluations of student behavioral skills. In related work, Koedel (2008) estimates high school teacher effects on graduation. However, he does not measure effects on non-cognitive skills and does not differentiate between effects that are due to improved cognitive skills versus non-cognitive skills. Finally, Mihaly, et. al. (2013) estimate teacher effects on non-test score outcomes to better predict teacher effects on test scores. However, they do not investigate whether their estimates capture teacher effects on economically meaningful outcomes that are not already captured by test scores.

measure only some dimensions of ability. If teachers improve skills not captured by test scores, then excellent teachers who improve long-run outcomes may not raise test scores, and the ability to raise test scores may not be the best predictor of effects on long-run outcomes. Indeed, Chetty, Friedman, and Rockoff (2011) note that teachers may have important effects on longer-run outcomes that are not reflected in their test score value-added and, using their same data, Chamberlain (2013) finds that test score effects may account for less than one quarter of the overall effect of teachers on college entry. This paper speaks to this second critique by being the first to investigate (a) whether teachers affect skills not captured by test scores, and (b) whether, and to what extent, teacher effects on a proxy for non-cognitive skills predict effects on long-run outcomes (that are missed by effects on test scores).

This paper is organized into four sections. The first section presents a latent factor model following Heckman, Stixrud, & Urzua (2006) in which both student and teacher ability have cognitive and non-cognitive dimensions. It shows that teacher effects on multiple short-run outcomes can predict effects on the same long-run outcome—even if the effects on the short-run outcomes are not correlated with each other. It also illustrates that the ability to predict variability in teacher effects on long-run outcomes will be greater with a combination of cognitive and non-cognitive outcomes than with any single outcome. The second section tests whether absences, suspensions, course grades, and on-time grade progression (in 9th grade) predict high school dropout and graduation, conditional on test scores. The third section estimates 9th grade Algebra and English teacher effects on test scores and non-test score outcomes. The fourth section tests the model and investigates the extent to which teacher effects on non-test score outcomes predict effects on high school completion (and proxies for college going) above and beyond that predicted by their test score effects alone.

The results from the second section show that most of the variability in absences, suspensions, grades, and grade progression is uncorrelated with test scores. Consistent with this, an underlying non-cognitive factor (i.e. a weighted average of these non-test score outcomes) explains covariance across these non-test score outcomes and is weakly correlated with test scores. This non-cognitive factor is associated with less high school dropout, increased high school graduation, and increased intentions to attend a 4-year college, all conditional on test scores. In survey data this non-cognitive factor also predicts fewer arrests, greater employment, and higher earnings, conditional on test scores — suggesting that the estimated non-cognitive factor is a proxy

for dimensions of ability not well measured by test scores.

In administrative data, 9th grade Algebra and English teachers have meaningful effects on test scores and non-test score outcomes. To address problems associated with student tracking and selection in secondary school, this paper follows Jackson (2014) and conditions on a student's school-track and also limits the analysis to a sub-sample of schools within which there is no selection of students to teachers based on observable characteristics. Also, quasi-experimental tests proposed in Chetty et al. (2011) are employed and suggest no bias. Interestingly, teacher effects on test scores and the non-cognitive factor are weakly correlated, so that many teachers that increase the non-cognitive factor do not raise test scores and *vice versa*. At the same time, teacher effects on both test scores and the non-cognitive factor predict effects on high school completion and proxies for college-going. Including teacher effects on the non-test score outcomes increases the predictable teacher-level variability in high school completion by 20 percent for Algebra teachers and over 200 percent for English teachers.

These results provide explanations for three seemingly conflicting findings in the recent literature. First, the importance of teacher effects on non-cognitive skills helps explain the Chamberlain (2013) finding that the effects of test score value-added on college-going presented in Chetty et al. (2011) reflect less than one-fifth of the total effect of teachers. Second, the relative importance of non-cognitive effects for English teachers can explain the Chetty et al. (2011) finding that English teachers have larger effects on long-run outcomes than math teachers despite smaller test score effects. Finally, the importance of non-cognitive skills offers a potential explanation for interventions with test score effects that "fade out" over time but have lasting effects on adult outcomes (Heckman et. al. 2013; Cascio & Staiger, 2012). More generally, these findings are the first to demonstrate that non-test-score outcomes can identify teachers who improve longer-run outcomes but are no more effective than average at improving test scores — findings that have direct implications for policy.

This paper is organized as follows: Section II presents the theoretical framework. Section III presents the data and relationships between long- and short-run outcomes. Section IV presents the empirical framework for obtaining causal estimates of teachers. Section V analyzes short-run teacher effects. Section VI analyzes how short-run teacher effects predict longer-run teacher effects and discusses the implications for policy. Section VII concludes.

II Theoretical Framework

This section presents a latent factor model following Heckman, Stixrud, & Urzua (2006) that justifies the use of *both* cognitive and non-cognitive outcomes to measure overall teacher quality. While students possess many types of cognitive and non-cognitive skills, the key insights from the model come from moving from a single to a multidimensional model of student ability. As such, for the sake of clarity, the model assumes only two broad ability types.

Student ability: Student ability is two-dimensional. One dimension is cognitive skill, and the other is non-cognitive skill. Each student *i* has ability vector $\nu_i = (\nu_{c,i}, \nu_{n,i})$, where the subscript *c* denotes the cognitive dimension and the subscript *n* denotes the non-cognitive dimension.

Teacher ability: Each teacher *j* has a two-dimensional ability vector, $\omega_j = (\omega_{c,j}, \omega_{n,j})$ where $E[\omega] = (0,0)$, which describes how much teacher *j* affects each dimension (cognitive or non-cognitive) of student ability. The total ability of student *i* with teacher *j* is thus $\alpha_{ij} = \nu_i + \omega_j$.

Outcomes: There are multiple outcomes y_z for each student *i*. Each outcome *z* is a linear function of the ability vector so that $y_{zij} = (\nu_i + \omega_j)'\beta_z$ where $\beta_z = (\beta_{c,z}, \beta_{n,z})$ is a vector of weights capturing the fact that some outcomes depend on cognitive ability (such as test scores) while others may depend on non-cognitive skills (such as attendance). There is an unobserved long-run outcome $y_{*ij} = \alpha_{ij}'\beta_* + \varepsilon_{*ij}$, where ε_{*ij} is random error and $\beta_{c,*}\beta_{n,*} \neq 0$. No two outcomes have the same relative weights on cognitive and non-cognitive ability. In the factor model representation, the two factors are the *total* ability of student *i* with teacher *j* in cognitive and non-cognitive ability, and vector β_z is the factor loadings for student outcome *z*.

Teacher Effects: The difference in student outcomes between teacher *j* with $\omega_j = (\omega_{c,j}, \omega_{n,j})$ and an average teacher with $\omega = (0,0)$ is a measure of *j*'s effect, relative to an average teacher. Teacher *j*'s effect for outcome *z* is therefore $\theta_{zj} = \omega_j'\beta_z$, so that teachers affect outcomes only through their effects on students' total ability. The long-run outcome is not observed, and policy-makers wish to predict teacher effects for long-run outcome $\theta_* = \omega_j'\beta_*$.

Proposition 1: *Teacher effects on long-run outcomes can be correlated with effects on multiple short-run outcomes even if effects on these short-run outcomes are uncorrelated with each other.*

Consider a case with two outcomes: y_1 and y_2. Suppose each outcome reflects only one

dimension of ability so that $\theta_{1j} = \beta_{c,1}\omega_{c,j}$ and $\theta_{2j} = \beta_{n,2}\omega_{n,j}$ where $\beta_{c,1}\beta_{n,2} \neq 0$. The two dimensions of teacher ability are uncorrelated, so $\text{cov}(\omega_{c,j}, \omega_{n,j}) = 0$. In this scenario, the covariance between teacher effects across all three outcomes are given by [1] through [3] below.

$$\text{cov}(\theta_1, \theta_2) = \text{cov}(\beta_{c,1}\omega_{c,j}, \beta_{n,2}\omega_{n,j}) = \beta_{c,1}\beta_{n,2}\,\text{cov}(\omega_{c,j}, \omega_{n,j}) = 0 \quad [1]$$

$$\text{cov}(\theta_1, \theta_*) = \text{cov}(\beta_{c,1}\omega_{c,j}, \beta_{c,*}\omega_{c,j}) + \text{cov}(\beta_{c,1}\omega_{c,j}, \beta_{n,*}\omega_{n,j}) = \beta_{c,1}\beta_{c,*}\,\text{var}(\omega_{c,j}) \neq 0 \quad [2]$$

$$\text{cov}(\theta_2, \theta_*) = \text{cov}(\beta_{n,2}\omega_{n,j}, \beta_{c,*}\omega_{c,j}) + \text{cov}(\beta_{n,2}\omega_{n,j}, \beta_{n,*}\omega_{n,j}) = \beta_{n,2}\beta_{n,*}\,\text{var}(\omega_{n,j}) \neq 0 \quad [3]$$

This illustrates that where student ability is multidimensional, both those teachers who improve cognitive ability (reflected in test scores) and those teachers who improve social skills (reflected in other outcomes) may improve long run outcomes (such as college attendance), even if these are different teachers. As such, teachers who improve outcomes not associated with test score gains may have important effects on longer-run outcomes. Section VI presents evidence of this.

Proposition 2: *One can predict a greater fraction of the variability in teacher effects on long-run outcomes using two short-run outcomes that reflect a different mix of both ability types than using any single short-run outcome.*

The best linear unbiased prediction of the teacher effect on the long-run outcome based on the effect on a single short-run outcome y_1 is the linear projection of effects on y_1 on the teacher's effect on the long-run outcome. Formally, $E[\theta_{*j} \mid \theta_{1j}] = \gamma\theta_{1j}$, where $\gamma = \text{cov}(\theta_*, \theta_1) / \text{var}(\theta_1)$.[4] The effect on the long-run outcome unexplained by θ_{1j} is $\ddot{\theta}_{*j} = (\beta_{c,*} - \gamma\beta_{c,1})\omega_{c,j} + (\beta_{n,*} - \gamma\beta_{n,1})\omega_{n,j}$. Consider another short-run outcome, y_2. The portion of θ_{2j} unexplained by θ_{1j} is $\ddot{\theta}_{2j} = (\beta_{c,2} - \pi\beta_{c,1})\omega_{c,j} + (\beta_{n,2} - \pi\beta_{n,1})\omega_{n,j}$ where $\pi = \text{cov}(\theta_2, \theta_1) / \text{var}(\theta_1)$. Teacher effects on additional outcome y_2 will increase the explained variability in teacher effects on the long-run outcome if $cor(\ddot{\theta}_{*j}, \ddot{\theta}_{2j}) \neq 0$. Because both residual effects $\ddot{\theta}_{*j}$ and $\ddot{\theta}_{2j}$ are linear functions of the same teacher ability, vector $\omega = (\omega_c, \omega_n)$, and linear functions of the same vector are generally correlated, it follows that in the vast majority of cases $cor(\ddot{\theta}_{*j}, \ddot{\theta}_{2j}) \neq 0$.

The model illustrates that with multidimensional ability, there may be improvements in our

[4] Note that $\theta_{*j} = \gamma\theta_{1j} + (\beta_{c,*} - \gamma\beta_{c,1})\omega_{c,j} + (\beta_{n,*} - \gamma\beta_{n,1})\omega_{n,j}$.

ability to predict teacher effects on long-run outcomes by evaluating teacher effects on multiple outcomes that reflect a variety of skills (rather than a single outcome).[5] Intuitively, with unidimensional ability, a second outcome does not improve our ability to predict effects on the long-run outcome because residual variability is random noise. However, with teacher effects through both cognitive and non-cognitive ability, residual variability in the effect on the long-run outcome may reflect dimensions of ability not captured by the first outcome. If the second outcome reflects different abilities from the first, the second outcome may explain residual variation in the effect on the long-run outcome. Section VI presents evidence of this.

III Data and Relationships between Variables

To estimate the effect of teachers on student outcomes, this paper uses data on all public school students in 9th grade in North Carolina from 2005 to 2011 from the North Carolina Education Research Data Center (NCERDC). The data include demographics, transcript data on all courses taken, middle school test scores, end of course scores for Algebra I and English I, and codes allowing one to link students' end of course test score data to individual teachers who administered the test.[6] I limit the analysis to students who took either the Algebra I or English I course (the two courses for which standardized tests have been consistently administered over time). Over 90 percent of all 9th graders take at least one of these courses, so the resulting sample is representative of 9th grade students as a whole. To avoid any bias that would result from teachers having an effect on students repeating 9th grade, the master data is based on the first observation for when a student is in 9th grade. Summary statistics are presented in Table 1.

These data cover 464,502 9th grade students in 619 secondary schools, 4,820 English I teachers, and 4,432 Algebra I teachers. The gender split is roughly even. About 58 percent of the 9th graders are white, 25.9 percent are black, 6.8 percent are Hispanic, two percent are Asian, and the remaining one percent are Native American, mixed race, or other. Regarding the highest education level of students' parents (i.e., the highest level of education obtained by either of the

[5] Note that this could also be true if it the case that both short-run outcomes were measured with error in a unidimensional model and also that outcomes 1 and 2 both measure the same dimension of ability. In this case the coefficient on the effect of outcome 1 in predicting the effect on the long-run outcome will be attenuated toward zero such that there may be some residual ability in the error term that could be picked up by the teacher effect on outcome 2. In section VI, I demonstrate that this is unlikely to be the case for the outcomes used in this paper.

[6] Because the teacher identifier listed is not always the student's teacher, I use an algorithm to ensure high quality matching of students to teachers. I detail this in Appendix note 1.

student's two parents), 6.5 percent were below high school, 39 percent had a high school degree, 15 percent had a junior college or trade school degree, 23.5 percent had a four-year college degree or greater, and 7.2 percent had an advanced degree (8 percent are missing data on parental education). The test score variables have been standardized to be mean zero with unit variance for each testing year. Incoming 7th and 8th grade test scores of students in the sample are about 25 percent of a standard deviation higher than that of the average in 7th and 8th grade. This is for two reasons; first, I only keep scores for the last time a student was in 7th or 8th grade. As such, the first score attempt for individuals with very low scores who also repeated 7th or 8th grade are not included in the analytic sample. Second, because the sample is of first time 9th graders, students with very low scores who repeated grades and may have dropped out of school before 9th grade are not included in this analytic sample.[7]

The main longer-run outcomes analyzed are measures of high school completion. Data on high school dropout and high school graduation (available through 2012) are linked to the 9th grade cohorts. Because students typically drop out of school before 12th grade, dropout data are available in 2012 for cohorts that were in 9th grade between 2005 and 2010, while graduation data are only available in 2012 for the 9th grade cohort from 2005 through 2009. To supplement the high school completion outcomes, data are also collected on AP courses taken (2008 and 2009 cohorts), SAT taking (2006 through 2009 cohorts), and self-reported intentions to attend a four-year college upon graduation (2006 through 2009 cohorts). In the sample, roughly 4.6 percent of 9th graders subsequently dropped out of school, while 82.5 percent graduated from high school. The remaining 11 percent either transferred out of the North Carolina school system or remained in school beyond the expected graduation year. The average 9th grader took 0.75 AP courses by 12th grade, 40 percent took the SAT by 12th grade, and 40 percent have intentions to attend a four-year college.

Correlations among the short run outcomes

The correlations among the 9th grade outcomes reveal some interesting patterns. The first pattern is that test scores are relatively strongly correlated both with each other and with grade point average (correlation≈0.6) but are weakly correlated with other non-test score outcomes. Specifically, the correlations between the natural log of absences (note: 1 is added to absences before taking logs so that zeros are not dropped) is -0.156 for Algebra test scores and -0.097 for

[7] The Algebra I and English I scores are also slightly above zero. This is because the sample of classrooms that can be well matched to teachers have slightly higher performance than average.

English test scores, and the correlations between being suspended are about -0.13 for both Algebra and English test scores. While slightly higher, the correlation between on-time progression to 10th grade (i.e. being a 10th grader the following year) and test scores is only 0.29. This reveals that while students who tend to have better test score performance also tend to have better non-test score outcomes, the ability to predict non-test score outcomes based on test scores is relatively limited. Simply put, students who score well on standardized tests are not necessarily those who are well-adjusted, and many students who are not well-behaved score well on standardized tests. Indeed, Table 2 indicates that test scores predict less than five percent of the variability in absences and being suspended, less than 10 percent of the variability in on-time grade progression, and just over one-third of the variability in GPA. Because these outcomes are interesting in their own right, test scores may not measure *overall* educational well-being.

The second notable pattern is that many behavioral outcomes are more highly correlated with each other than with scores. For example, the correlations between suspensions and test scores are smaller than those between suspensions and all the other outcomes. Similarly, the correlations between absences and test scores are smaller than those between absences and the other outcomes. The third notable pattern is that GPA is relatively well correlated with both the test score and the non-test score outcomes. The fact that GPA is correlated with both test scores and non-test-score outcomes is consistent with research (e.g., Howley, Kusimo, & Parrott, 2000; Brookhart, 1993) finding that most teachers base their grading on some combination of student product (exam scores, final reports, etc.), student process (effort, class behavior, punctuality, etc.) and student progress — so that grades reflect a combination of cognitive and non-cognitive skills. In sum, in the context of the model, the patterns imply three groups of variables: those that are mostly cognitive (English I and Algebra I test scores), those that are mostly non-cognitive (absences and suspensions) and those that reflect a combination of cognitive and non-cognitive ability (on-time grade progression and GPA). If teachers improve student outcomes through improving both cognitive and non-cognitive skills, their effect on a combination of these abilities should better predict their effect on longer-run outcomes than test scores alone.

The relationship between short-run and longer-run outcomes

Much of the justification for the use of test scores to measure the effectiveness of educational interventions is that higher test scores predict improved adult outcomes. To make a similar case for also using non-cognitive outcomes, evidence is presented that (a) there is an

underlying non-cognitive factor that explains much of the covariance between non-test score outcomes and is only moderately correlated with test scores; and (b) increases in both test scores and this estimated non-cognitive factor are independently associated with better adult outcomes. This analysis is descriptive and may not represent causal relationships because there may be unaccounted for differences in incentives and contexts that generate different outcomes (Heckman & Kautz, 2012). However, Section VI presents the relationship between exogenous changes in 9th grade outcomes and longer-run outcomes that can be interpreted causally. The key longer-run outcomes studied are measures of high school completion.

Table 3 shows that both test scores and non-test-score outcomes independently predict long-run outcomes. I regress longer-run outcomes (from 12th grade) on GPA, absences, being suspended, on-time grade progression, and test scores (all measured in 9th grade). To remove the influence of differences in socioeconomic status or demographics, all models include controls for parental education, gender, and ethnicity, and include indicator variables for each secondary school. Columns 1 and 2 show that while higher test scores in 9th grade do predict less dropout and more high school graduation, the non-test score outcomes in 9th grade also predict variability in these important longer-run outcomes conditional on test scores. As one might expect, higher GPAs and on-time grade progression are associated with lower dropout rates and more high school graduation. Similarly, increased suspensions and absences are associated with increased dropouts and lower high school graduation. For both outcomes, one can reject the null hypotheses that the 9th grade non-test score outcomes have no predictive power for longer-run outcomes conditional on test scores at the one percent level.

To ease interpretation, I created a weighted average of the non-test score outcomes as a single proxy for non-cognitive skills. To do this, I estimated a factor model on the four non-test score outcomes (absences, suspensions, GPA, and on-time grade progression) and computed the unbiased prediction of the first underlying factor as my proxy for non-cognitive ability.[8] This average was then standardized to be mean zero unit variance. This weighted average is an estimate of the underlying ability that explains the covariance in these non-test score outcomes. Table 2 presents the fraction of the variability in outcomes explained by this factor. This factor explains 46 percent of the variability in absences, 24 percent of the variability in being suspended, 71

[8] This predicted factor was computed using the Bartlett method, however the results are robust to other methods. The predicted factor is Factor = -0.45*absenses -0.35*suspended +0.64*GPA +0.57*on time in 10th grade.

percent of the variability in GPA, and 40 percent of the variability in on-time grade progression. Because students with higher test scores tend to have better outcomes in general, this factor explains a modest 24 and 28 percent of the variability in Algebra and English test scores, respectively. In sum, this factor captures the common variability in the non-test-score outcomes and is moderately correlated with test scores.

Columns 3 and 4 of Table 3 show that for both longer-run outcomes a standard deviation (σ) increase in the non-cognitive factor is associated with larger improvements than a standard deviation increase in Algebra test scores (results are similar using English test scores). Specifically, while a 1σ increase in test scores is associated with a 1.3 percentage point decrease in dropout, a 1σ increase in the non-cognitive factor is associated with a 5.51 percentage point decrease in dropout. Similarly, while a 1σ increase in test scores is associated with a 2.66 percentage point increase in high school graduation, a 1σ increase in the non-cognitive factor is associated with a 15.7 percentage point increase. While the following are not the main longer-run outcomes, I analyze effects on the number of AP courses taken, whether a student takes the SAT by 12^{th} grade, and intentions to attend a four-year college at high school graduation (proxies for college-going). These variables are not available for most years (note the different sample sizes), but are analyzed to provide additional supporting evidence. As with dropout and graduation, a 1σ increase in the non-cognitive factor is associated with larger improvements in these additional longer-run outcomes than a 1σ increase in test scores (columns 5 through 7). Taken at face value, this suggests that the non-cognitive factor may be as good, if not better, a predictor of dropout, high school graduation, and college intentions as test scores. While these results are largely descriptive, these relationships are strikingly similar to those in Section VI based on exogenous changes in test scores and the non-cognitive factor caused by teachers.

To validate the use of the factor, I replicate the patterns in Table 3 using nationally representative data — the National Educational Longitudinal Survey of 1988 (NELS-88; see appendix note A3). As in the NCERDC data, for both dropout and high school graduation, a 1σ increase in the non-cognitive factor is associated with much larger effects than a 1σ increase in math scores in 8^{th} grade. Looking to other adult outcomes in the survey data, the non-cognitive factor predicts much variability in being arrested, working, and earnings (all at age 25), conditional on test scores (Table A3). Specifically, a 1σ increase in the non-cognitive factor is associated with being 4.54 percent less likely to be arrested (a 22 percent reduction relative to the sample mean),

15.3 percentage points more likely to be employed, and earning 20 percent more, conditional on test scores. To further validate the use of the factor, I look for similar behaviors between the factor and other measures of non-cognitive skills. Psychometric measures of non-cognitive skills have been found to be particularly important at the lower end of the earnings distribution (Lindqvist & Vestman, 2011; Heckman, Stixrud, & Urzua, 2006). To see if this is also true for the non-cognitive factor, I estimate the marginal effect of the factor on log earnings at different points in the earnings distribution using the NELS-88. Similar to psychometric measures of non-cognitive skills, the non-cognitive factor has much larger effects at the lower end of the earnings distribution — thereby suggesting that this factor is a reasonable proxy for non-cognitive ability.

While I am agnostic about the exact skills captured by this factor, low levels of agreeableness and high neuroticism are associated with more school absences, externalizing behaviors, juvenile delinquency, and lower educational attainment (Lounsbury, Steel, Loveland, & Gibson, 2004; Barbaranelli, Caprara, Rabasca, & Pastorelli, 2003; John, Caspi, Robins, Moffit, & Stouthamer-Loeber, 1994; Carneiro, Crawford, & Goodman, 2007). Also, high conscientiousness, persistence, grit, and self-regulation are all associated with fewer absences, fewer externalizing behaviors, higher grades, on-time grade progression, and higher educational attainment (Duckworth, Peterson, Matthews, & Kelly, 2007). This suggests that the factor reflects a skill-set associated with high conscientiousness, high agreeableness, and low neuroticism, and is correlated with self-regulatory skills, persistence, and grit. Irrespective of what we call it, the key point is that this non-cognitive factor captures skills that explain certain observable outcomes not explained by test scores and may predict long-run success.

The results show that the non-cognitive factor is a reasonable proxy for a dimension of soft or non-cognitive skills and explains variability in adult outcomes above and beyond that explained by test scores. In the context of the model, the patterns imply that (a) teachers who improve the non-cognitive factor may have effects on important long-run outcomes that may go undetected by their effects on test scores, and (b) evaluating a teacher's effects on both test scores and the non-cognitive factor might improve our ability to identify excellent teachers who improve student well-being overall by improving both cognitive and non-cognitive student ability. These predictions are tested directly in section VI.

IV Empirical Strategy

This section outlines the strategy used to estimate and predict teacher effects on student test-score outcomes and non-test-score outcomes in 9th grade. It then previews how these predictions can be used to identify those teachers that improve longer-run outcomes. Finally, it presents evidence that estimated and predicted teacher effects on short-run outcomes (test scores and non-test-score outcomes in 9th grade) are valid such that they can be used to predict teacher effects on longer-run outcomes (measures of high school completion).

To estimate teacher effects on short-run (9th grade) outcomes, I follow standard practice in the literature and model student outcomes as a function of lagged student achievement and other covariates. I model outcome Y_{icjgys} of student i in class c with teacher j in school track sg, at school s, in year y in equation [4] (note: most teachers are observed in multiple classes).

$$Y_{icjgys} = A_{iy\text{-}1}\delta + X_i\beta + I_{ij}\cdot\theta_j + I_{sgi}\,\theta_{sg} + I_{sy}\,\theta_{sy} + \phi_c + \varepsilon_{icjgys} \qquad [4]$$

$A_{iy\text{-}1}$ is a matrix of incoming achievement of student i (7th and 8th grade math and reading scores); X_i is a matrix of student-level covariates (parental education, ethnicity, and gender); I_{ij} is an indicator variable equal to 1 if student i has teacher j and equal to 0 otherwise so that θ_j is a time-invariant fixed effect for teacher j; I_{sy} is an indicator variable denoting whether the student is in school s in year y so that θ_{sy} is a school-by-year fixed effect; ϕ_c is a random classroom-level shock; and ε_{icjgys} is a random error term. The key conditioning variable is I_{sgi} which is an indicator variable equal to 1 if student i is in school track sg and 0 otherwise so that θ_{sg} is a time-invariant fixed effect for school track sg. Following Jackson (2014), a school track is the unique combination of the ten largest academic courses, the level of Algebra I taken, and the level of English I taken in a particular school.[9] As such, only students at the same school who take the same academic courses, level of English I, and level of Algebra I are in the same school track.[10] Because many students pursue the same course of study, less than one percent of all students are in singleton tracks, 80

[9] While there are many courses that 9th grade students can take (including special topics and reading groups), there are 10 academic courses that constitute two-thirds of all courses taken. They are listed in Appendix Table A1. Defining tracks flexibly at the school/course-group/course level allows for different schools that have different selection models and treatments for each track. Even though schools may not have explicit labels for tracks, most practice de-facto tracking by placing students of differing levels of perceived ability into distinct groups of courses (Sadker & Zittleman, 2006; Lucas & Berends, 2002). As highlighted in Jackson (2014) and Harris & Anderson (2012), it is not only the course that matters but also the levels at which students take a course.

[10] Students taking the same courses at different schools are in different school-tracks. Students at the same school in at least one different academic course are in different school tracks. Similarly, students at the same school taking the same courses but taking Algebra or English at different levels are in different school tracks.

percent of students are in tracks with more than 30 students, and the average student is in a school track with 179 other students.

The key identifying assumption for consistent estimation of teacher effects, θ_j, is that there is no selection of students to teachers within tracks.[11] Including indicators for each school track in a value-added model compares outcomes across teachers within groups of students *in the same track at the same school*. This removes the influence of both track-level treatments and selection to tracks on estimated teacher effects. Because the models include school-by-track effects, all inference is made within school tracks so that identification of teacher effects comes from two sources of variation: (1) comparisons of teachers at the same school teaching students in the same track *at different points in time* and (2) comparisons of teachers at the same school teaching students in the same track *at the same time*. The first source of variation is driven entirely by changes in staffing over time within schools (e.g., the Algebra I teacher in the advanced track at Hope High School was Ms. Smith in 2005 and Mr. Jones in 2006). This source of variation is valid as long as students do not select across cohorts within tracks (e.g., skip a grade) or schools in response to changes in Algebra I and English I teachers. Tests in section IV.2 show no evidence of such selection. The second source of variation comes from having multiple teachers for the same course in the same track at the same time (e.g., both Ms. Smith and Mr. Jones teach Algebra I in the advanced track at Hope High School in 2006). This source of variation is robust to student selection *to* school tracks and is valid as long as students do not select to teachers *within* school-track-year cells. Tests in section IV.2 show that the findings are not driven by student selection within school-track-years.[12]

IV.1 Estimating Predicted Teacher Effects

The key objective of this study is to determine whether teachers who improve test scores or non-test-score outcomes also improve longer-run student outcomes. To do this, one must compare longer-run outcomes of students who are exposed to teachers with different estimated

[11] In these models, the teacher effects are teacher-level means of the outcome after adjusting for incoming student characteristics, school-by-year level shocks, and school-by-track effects. For test score outcomes, this model is a standard value-added model with covariate adjustments.

[12] To compare variation within school tracks during the same year to variation within school tracks across years (cohorts), I computed the number of teachers in each non-singleton school-track-year-cell for both Algebra I and English I (Appendix Table A2). About 63 and 51 percent of all school-track-year cells include one teacher in English I and Algebra I, respectively. As such, much variation is likely based on comparing single teachers across cohorts within the same school track. Section V.2 shows that results using variation within school-track-cohort cells are similar to those obtained using only variation within school tracks but across cohorts.

effects on short-run outcomes. However, because classes that have better test scores for reasons unrelated to the teachers tend to have higher high school graduation rates for those same reasons, it is important that the estimated short-run effects are not based on the same students for whom the longer-run effects are being examined. Doing so would lead to a mechanical endogeneity. To deal with this issue, I follow a strategy very similar to Chetty et. al. (2011) to form a prediction of how much each teacher will improve her students' test scores or non-test score outcomes in a given year based on her performance in all *other* years (based on a different set of students). Comparing the longer-run outcomes of students exposed to teachers with different predicted effects on short-run outcomes based on data from all other years removes the mechanical endogeneity that would emerge due to common shocks and isolates variability in a teacher's predicted effect that is persistent over time.

This prediction based on other years of data is computed in two steps:

<u>*Step 1:*</u> Estimate equation [5] where data for year *y* is excluded from the estimation sample when creating a prediction of teacher effects for year *y*.

$$Y_{icjgys} = A_{iy\text{-}1}\delta + X_i\beta + I_{sgi}\,\theta_{sg} + I_{sy}\,\theta_{sy} + \phi_c + \theta_j + \varepsilon_{icjgys} \qquad [5]$$

There are no teacher (or classroom) indicator variables included so the total error term is $\varepsilon^{*}=\phi_c + \theta_j + \varepsilon_{igjy}$ (i.e., a teacher effect, a classroom effect, and the error term). I then compute, $\mu_{j,y'}$, the average student-level residual for each teacher.[13] Under the identifying assumptions, these "leave out means" are consistent estimates of teachers' effects on 9th grade outcomes in other years.

<u>*Step 2*</u>: In the second step, I obtain the *predicted* effect of a teacher for the current year based on the estimate of her effect in all other years computed in Step 1. To do this, I estimate equation [6] where $\mu_{j,y'}$ is the out-of-sample teacher-level mean residual (from other years).

$$Y_{icjgys} = A_{iy\text{-}1}\delta + X_i\beta + \gamma\mu_{j,y'} + I_{sgi}\,\theta_{sg} + I_{sy}\,\theta_{sy} + \phi_c + \varepsilon_{icjgys} \qquad [6]$$

The prediction of each teacher's effect from [6] is $\hat{\gamma}\mu_{j,y'}$, and is the out-of-sample teacher-level mean residual multiplied by its coefficient in a regression predicting that same outcome in-sample. If performance in the past were a prefect predictor of current performance, then $\hat{\gamma} = 1$ and the out-of-sample estimate would be the predicted effect. However, because the out-of-sample estimates are estimated with error, $\hat{\gamma} < 1$, so that the prediction "shrinks" the out-of-sample estimate toward zero. When the prediction ($\hat{\gamma}\mu_{j,y'}$) based on test scores (or the non-cognitive factor) is included as

[13] I use the precision weighted mean as proposed by Kane & Staiger (2008). See appendix note 2 for details.

a regressor for longer-run outcomes, the coefficient on $\hat{\gamma}\mu_{j,y'}$ represents the marginal effect of having a teacher that is predicted to increase test scores (or the non-cognitive factor) by one standard deviation. By construction, the coefficient of a teacher's predicted effect on test scores (or the non-cognitive factor) on test scores (or the non-cognitive factor) is equal to one.

Once these predictions of the teacher's effect on test scores and teacher's effect on the non-cognitive factor are obtained, one can test whether teachers who improve test scores or the non-cognitive factor also improve longer-run outcomes using a regression predicting the longer-run outcomes on the predicted short-run effects. I expand on this approach in Section VI.

The variance of the estimated teacher effects $\hat{\theta}_j$ from [4] will overstate the true variance of teacher quality because of sampling variation and classroom shocks. Also, the estimated variance of $\hat{\gamma}\mu_{j,y'}$ from [6] will understate the effect of teachers because the predictions are shrunk toward zero (Chetty et. al., 2013a). As such, I follow Kane and Staiger (2008) and use the covariance between mean classroom-level residuals for the same teacher as a measure of the variance of teacher effects.[14] Following Jackson (2014), I compute bootstraped standard errors for the estimated covariance and use them for normal-distribution-based confidence intervals.[15]

IV.2 Addressing Selection of Students to Teachers

Before predicting teacher effects on longer-run outcomes based on their performance at improving test scores and the non-cognitive factor, it is important to ensure that within the estimation sample there is no selection of students to teachers within school-tracks. While many studies rely on the assumption that teachers are randomly assigned to students conditional on incoming test scores (Koedel & Betts, 2011; Kinsler, 2012; Kane & Staiger, 2008), the key identifying assumption in this paper is that teachers are randomly assigned to students within

[14] I then compute mean residuals from [5] for each classroom. Then I link every classroom-level mean residual and pair it with another random classroom-level mean residual for the same teacher and compute the covariance of these mean residuals. If the classroom errors ϕ_c are uncorrelated with each other and uncorrelated with teacher quality θ_j, the covariance of mean residuals within teachers but across classrooms is a consistent measure of the true variance of persistent teacher quality ($cov(\bar{e}^*_c, \bar{e}^*_{c'} \mid J = j) = \text{cov}(\theta_j, \theta_j) = \text{var}(\theta_j) \longrightarrow \sigma^2_{\theta_j}$). I replicate this calculation 100 times and take the median of the estimated covariance as the parameter estimate.

[15] I use the standard deviation of 100 randomly computed "placebo" covariances (i.e., sample covariances across classrooms for different teachers) to form an estimate of the standard deviation of the sampling distribution of the covariance across classrooms for the *same* teacher.

school tracks (with no need for additional controls or lagged test scores).[16] In other words, while other studies rely on conditioning on lagged test scores to remove bias due to selection, this study relies on conditioning on school tracks to remove bias due to selection. This section describes how selection on observables is addressed and then presents a test for selection on unobservables.

IV.2.a *Addressing Selection on Observables*

To assess the degree to which there is selection on observables within school tracks, I predict each outcome (based on a linear regression of each outcome on 7th and 8th grade math and reading scores, parental education, gender, and ethnicity) and then regress predicted student outcomes (an index of all observable incoming student characteristics) on predicted teacher effects (estimated out-of-sample) while controlling for school track indicators and year indicators. If there were no selection of students to teachers within school tracks, there would be no systematic relationship between predicted outcomes and predicted teacher effects.

When the aforementioned test for selection on observables is implemented on the full sample of schools, there is evidence of positive selection for test score value-added for both subjects, on average. That is, for both subjects, students with better observable characteristics tend to be assigned to teachers that are predicted to improve test scores (there is no such selection for teacher effects on the non-cognitive factor). Because the set of controls included are strong predictors of student outcomes (parental education, incoming test scores, gender, ethnicity) this does not imply that there is bias because the observable covariates might capture most of the important sources of selection. However, because there is no way to know if the observable covariates are sufficient to control for all selection, when there is selection on observables there is always the worry that there may also be selection on unobserved dimensions (Altonji, Elder, & Taber, 2005). If there were no selection on observables, it would be plausible that there is no selection on unobserved dimensions. The approach taken in this paper is therefore to obtain a subsample of high schools within which there is no selection on observables.

If all schools had positive selection to teachers there would be no way to obtain a sample of schools within which the assumption of no selection is satisfied. However, if some schools have positive selection to teachers while others do not, one can remove those schools that exhibit strong sorting to obtain a subsample of schools within which there is no sorting of students to teachers on

[16] The tests presented indicate that, within the schools in the preferred sample, conditioning on tracks is sufficient to remove selection bias (without having to condition on lagged test scores).

observables. Because all inferences are based on within-school comparisons, one can obtain *internally* valid inferences within this subsample of schools. I create this subsample in two steps. First, I regress predicted outcomes on predicted teacher effects for each school and save the t-statistic associated with the null hypothesis that there is no relationship for each school (this is a measure of the degree of within-school sorting for each school). Second, to remove schools with strong sorting, I remove those schools with t-statistics above some maximum. To obtain the final "clean" sample, I chose the maximum t-statistic such that the coefficient on the predicted teacher effects are zero for both teacher effects on test scores and the non-cognitive factor. See appendix Figure A1 for a plot of the maximum t-statistic and the degree of sorting on observables on English and Algebra test score value-added. Removing schools with strong sorting to achieve no sorting on observables leaves 70 percent of the English I sample and 85 percent of the Algebra I sample. While this does limit the ability to speak about teachers at all North Carolina schools, all inferences are based on within-school comparisons so all inferences will be internally valid. To assuage any lingering concerns, Appendix Note 4 presents a Monte Carlo simulation showing that this procedure yields unbiased teacher effect estimates under selection on observables.

The top panel of Table 4 shows the result of this sample restriction. The coefficient on the predicted teacher effect on Algebra scores on predicted Algebra scores is 0.0095 (se=0.0439) and that for the Algebra teacher effects on the non-cognitive factor on the predicted non-cognitive factor is 0.0204 (se=0.272). That is, within the subsample of schools that do not exhibit strong selection of students to teachers, a teacher who is predicted to increase Algebra scores by one standard deviation increases predicted scores by less than one percent of a standard deviation, and a teacher who is predicted to increase the non-cognitive factor by one standard deviation increases the predicted factor by two percent of a standard deviation (neither relationship is statistically significantly different from zero). For English, the patterns are similar. Within the subsample of schools that do not exhibit strong selection of students to teachers, a teacher who is predicted to increase English scores by one standard deviation increases predicted scores by less than one percent of a standard deviation, and a teacher who is predicted to increase the non-cognitive factor by one standard deviation *decreases* the predicted factor by six percent of a standard deviation (neither relationship is statistically significantly different from zero).

In sum, within the restricted subsample there is no tendency for better students to be assigned to better or worst teachers on any observable dimension. Because all estimates are

obtained by comparing student outcomes across teachers within schools, removing entire schools that exhibit selection within tracks does not introduce bias or endogeneity. Because there is no selection of students to teachers based on a rich set of observable covariates in this subsample, it is plausible that there is no student selection to teachers in unobserved dimensions within this same subsample. Section IV.2.b presents a test indicating that this is likely the case.

IV.2.b *Selection on Unobservables*

Having imposed the condition of no selection on observables, I now turn to selection on unobservables. To test for selection on unobservables within school track cohorts, I follow Chetty, Friedman, and Rockoff (2011) and exploit the statistical fact that the effects of any selection among students within the same school track and cohort will be eliminated by aggregating the treatment to the school-track-year level and relying only on cohort-level variation across years within school tracks. That is, if the estimated teacher effects merely capture student selection to teachers within school track cohorts, then the arrival of a teacher with a high positive predicted effect (who increases the average predicted teacher effect for a cohort but has no effect on real teacher quality or student outcomes) should have no effect on average student outcomes for that cohort. Conversely, if the predicted effects are real, differences in average predicted teacher quality across cohorts (driven by changes in teaching personnel within schools over time) should be associated with similar differences across cohorts in average cohort-level outcomes as the same difference in estimated teacher quality across individual students (due to there being multiple teachers in the same school track at the same time) within the same cohort.

To test this, I estimate equations [7] and [8], where $\hat{\theta}_j$ is the predicted (out-of-sample) effect of teacher *j*, $\overline{\hat{\theta}}_{j \in sgy}$ is the mean predicted teacher effect in school track s*g* in year *y*, θ_{sg} is a school track effect, θ_{sy} is a school-year effect, and θ_{sgy} is a school-track-year effect.

$$Y_{isgjcy} = A_{iy-1}\delta + \psi_1\hat{\theta}_j + X_{iy}\beta + \theta_{\text{sgy}} + \varepsilon_{isgjcy} \quad [7]$$

$$Y_{isgjcy} = A_{iy-1}\delta + \psi_2\overline{\hat{\theta}}_{j \in sgy} + X_{iy}\beta + I_{sgi}\,\theta_{\text{sg}} + \theta_{\text{sy}} + \varepsilon_{isgjcy} \quad [8]$$

Equations [7] and [8] both predict outcomes as a function of estimated teacher effects on student outcomes, but each uses a distinct source of variation. In equation [7], teacher quality is defined at the student level. The model includes a track-school-year fixed effect, so that it only makes comparisons among students with different teachers in the same school track and year (removing all variation due to personnel changes over time). In contrast, by defining teacher quality at the school-track-cohort level in equation [8], one no longer compares students within the same school-

track-year (where selection is likely) and only compares entire cohorts of students in the same school track over time (where selection is unlikely, because variation in this aggregate measure is due only to changes in the identities of teachers in the school track over time). To control for school-level changes that could affect the cohort-level results, all models include school-by-year fixed effects. Standard errors are adjusted for clustering at the teacher level in equation [7] and the school track level in equation [8]. Relating the predictions to the equations directly, if there is no sorting ψ_1 should be similar to ψ_2, and if the effects are due to sorting then ψ_2 will be equal to 0. Note that because the predicted effects are scaled to have a coefficient of 1 in the preferred specification (that uses all the variation), where there is little bias the coefficients for both ψ_1 and ψ_2 should be close to 1.

The results are presented in the lower panel of Table 4. First I look at test score effects. Despite there being no relationship between predicted teacher quality and predicted outcomes, there are economically and statistically significant effects of predicted teacher quality on *actual* outcomes for both subjects. Using all the variation (preferred model) the coefficient on the predicted test score effect on actual test scores is 1 (by construction) for both Algebra and English teachers (p-value<0.01 for both subjects). Using variation only within school-track-cohorts, the coefficients for predicted test score effects on actual test scores are 0.981 and 1.03 for Algebra and English, respectively. Using only variation across school track cohorts, the coefficients on the predicted test score effects on actual test scores are 1.051 and 0.898 for Algebra and English, respectively. For neither the within- nor across-track models is the coefficient on predicted teacher quality statistically distinguishable from 1, suggesting no selection on unobservables.

I now turn to effects of the predicted teacher effects on the non-cognitive factor. By construction, the coefficient for a teacher's predicted effect on the non-cognitive factor is 1 for both Algebra and English teachers in the preferred model. However, there are large differences in precision across the two subjects. While the out-of-sample estimate is statistically significantly different from zero at the one percent level for English teachers, it is only significant at the 10 percent level for Algebra teachers. The estimates indicate that with 95 percent confidence an English teacher that is predicted to improve the non-cognitive factor by 1σ can be expected to improve the non-cognitive factor by between 0.57 and 1.42σ. In contrast, with 95 percent confidence an Algebra teacher that is predicted to improve the non-cognitive factor by 1σ can be expected to improve the non-cognitive factor by between -0.2 and 2.2σ. Simply put, the ability to

predict a teacher's effect on the non-cognitive factor based on her performance in other years is good for English teachers and limited for Algebra teachers.

Looking specifically at selection on unobservables, the within-track estimates are 1.024 and 0.84 for Algebra and English, respectively. The across-track estimates of mean predicted teacher quality on the non-cognitive factor are 0.911 and 1.524 for Algebra and English, respectively. As with test score effects, the within-track and across-track estimates are not statistically significantly different from one – indicating no selection on unobservables. While there is no evidence of selection, the effect of Algebra teachers on the non-cognitive factor is not statistically significantly different from zero. This suggests that while Algebra teacher effects on the non-cognitive factor may be unbiased, they do not allow for precise out-of-sample predictions. This will likely limit the ability to predict effects on longer-run outcomes using Algebra teacher effects on the non-cognitive factor. I show evidence of this in Section VI.

The discussion thus far has focused on selection within school tracks. However, one might wonder if the results are biased due to student selection across tracks. To test this, I regress student outcomes on the school-year level mean predicted teacher effects. If the results are driven by student selection across tracks, then the school-year average effects (aggregated across school tracks) should have no effect on outcomes. Also, if the estimated effects are not driven by selection across tracks, the estimates based on the school-level mean effects should be similar to those for the individual teacher effects. The result in Table 4 show that mean school cohort level teacher quality have effects on all outcomes similar to those from teacher-level variation – indicating that selection across tracks does not bias the results. Having created a sample within which there is no selection on observables, and established the there is no selection to teachers on unobservables, I now analyze teacher effects using the clean analytic samples.

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

V.1 True Variance of Teacher Effect on Test Score and Non-Test Score Outcomes

Before presenting the effects of being assigned to a teacher with higher predicted effects on 9th grade outcomes, I examine the magnitudes of the teacher-level variability that is persistent across classrooms for each of the short-run outcomes. Table 5 presents the square root of the estimated covariance across classrooms for the same teachers. I also present the 95 percent confidence intervals for the estimated standard deviations. In the few instances where the sample

covariance is negative, I report the standard deviation as zero. Note if the true covariance is zero, one will obtain negative covariance half of the time. Table 5 presents the results from two specifications. The top panel presents results from the preferred model that includes both school track fixed effects and school-by-year fixed effects to account for both bias due to tracking and any school-wide shocks that may confound teacher effects. To illustrate the importance of accounting for school tracks and transitory school-level shocks, the lower panel presents results from a simple model with school fixed effects, year fixed effects, and student controls.

In models that only include school fixed effects and year fixed effects, both Algebra and English teachers have sizable and economically meaningful effects on test scores and non-test-score outcomes. Also, in such models, English teachers have large statistically significant effects on Algebra test scores and Algebra teachers have marginally statistically significant effects on English test scores. In the preferred models that include school-by-year effects and school-track effects, the variability of all the effects fall by between one half and two thirds – indicating that accounting for track-level variability and transitory school shocks is important (Jackson, 2014). In the preferred models, there are no statistically significant effects across subjects. That is, the 95 percent confidence intervals for Algebra teacher effects on English test scores and the confidence intervals for English teacher effects on Algebra test scores both include zero. While it is possible that there could be non-zero cross-subject effects, the fact that there are no cross-subject effects despite statistically significant own-subject effects lends credibility to the empirical design. Accordingly, I focus the remainder of the discussion on this preferred model.

In the preferred model (top panel for each subject), the standard deviation of the Algebra teacher effects on Algebra test scores is 0.0656σ and one can be 95 percent confident that the true standard deviation is between 0.0531σ and 0.0761σ. While the estimated standard deviations of Algebra teacher effects on the non-test-score outcomes are all positive, the 95 percent confidence intervals for effects on suspensions, absences, and on-time grade progression all include zero. However, the standard deviation of the effect on GPA is 0.0368 grade points, and this is statistically significantly different from zero. The confidence interval implies that having an Algebra teacher at the 85th percentile of effects on GPA versus the 15th percentile would be associated with between 0.02 and 0.12 grade points higher GPA. Looking to the non-cognitive factor that combines all these non-test score outcomes into a single variable, the standard deviation of Algebra teacher effects on the factor is 0.0725σ, where the 95 percent confidence interval is

between 0.0513σ and 0.088σ. The more precise effects on the non-cognitive factor likely reflect the facts that the weighted average of four outcomes will have less measurement error than any one individual outcome. Also, the fact teacher-level variability is statistically significant for only one of the four non-test score outcomes may explain why Algebra teacher effects on the non-cognitive factor led to imprecise out-of-sample predictions in Table 4.

Looking to English teachers, in the preferred model the standard deviation of English teacher effects on English test scores is 0.0298σ, and one can be 95 percent confident that the true standard deviation is between 0.0155σ and 0.0392σ. Unlike for Algebra teachers, the estimated standard deviations of English teacher effects are statistically different from zero for being suspended, GPA, and on-time 10^{th} grade enrollment: the standard deviation of teacher effects on GPA is 0.0418 grade points, that on suspensions is 0.152, and that on enrolling in 10^{th} grade is 1.66 percentage points. To put these estimates into perspective, having an English teacher at the 85^{th} percentile of effects on GPA versus the 15^{th} percentile would be associated with 0.084 grade points higher GPA, being three percentage points less likely to have been suspended, and being 3.2 percentage points (0.1σ) more likely to enroll in 10^{th} grade on time. Summarizing these effects, the standard deviation of English teacher effects on the non-cognitive factor is 0.0648σ, and one can be 95 percent confident that the true standard deviation is between 0.0515σ and 0.0758σ.

Overall, for both subjects, having a teacher at the 85^{th} percentile of improving non-cognitive ability versus the 15^{th} percentile would be associated with between 0.12σ and 0.15σ higher non-cognitive ability. However, a few key differences are worth noting across subjects: Algebra teachers have similarly sized effects on Algebra test scores as they do on the non-cognitive factor, while English teachers have much larger effects on the non-cognitive factor than on English test scores. Given that the differences in longer-run outcomes associated with a 1σ increase in the non-cognitive factor are larger than that of a 1σ increase in test scores (Table 3), this implies that including teacher effects on the non-cognitive factor would have a large effect on our ability to predict teacher effects on longer-run outcomes, particularly for English teachers. The results in section VI show that this is the case.

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

Having established that teachers have real causal effects on test scores and non-test score outcomes, this section documents the relationships between these estimated effects. To gain a sense of whether teachers who improve test scores also improve other outcomes, I regress the

predicted teacher effects for all the outcomes on the effects on Algebra test scores, English test scores, and the non-cognitive factor. The reported R^2s in Table 6 measure the fraction of the variability in the predicted teacher effect on each outcome that can be explained by (or is associated with) teacher effects on test scores or the non-cognitive factor.

The top panel presents effects for Algebra teachers. Algebra teachers with higher test score effects are associated with better non-test score outcomes, but the relationships are weak. Effects on Algebra test scores explain less than one percent of the variance in teacher effects on suspensions, 1.6 percent for absences, 2.1 percent for GPA, and 5.7 percent of the effects on on-time 10th grade enrollment (top panel top row). This indicates that while teachers who raise test scores may also be associated with better non-test score outcomes, most of the effects on non-test score outcomes are unrelated to effects on test scores. It is worth noting that while student GPA and test scores are quite highly correlated across students (Table 2), variability in teacher effects on test scores predict little of the variability in their ability to raise GPA, and *vice versa*. As expected, effects on the non-cognitive factor explain much of the effects on the non-test-score outcomes. Specifically, Algebra teacher effects on the non-cognitive factor explain 34.8 percent of the estimated teacher effect on suspensions, 50 percent for absences, 63.5 percent for GPA, and 37.9 percent of the effect on on-time 10th grade enrollment (top panel second row). However, teacher effects on the non-cognitive factor explain only 5.7 percent of the variance in estimated teacher effects on Algebra scores. Results for English teachers (lower panel) are similar to those for Algebra teachers. English teacher effects on English test scores explain little of the estimated effects on non-test-score outcomes. Specifically, teacher effects on English test scores explain less than five percent of the variance of teacher effects on suspensions, absences, GPA, on-time 10th grade enrollment, and the non-cognitive factor (lower panel top row). In contrast (and as expected), English teacher effects on the non-cognitive factor explain 26.7 percent of the variance in teacher effects on suspensions, 43.9 percent for absences, 61.9 percent for GPA, and 43.4 percent for enrolling in 10th grade one year after 9th grade enrollment.

For both subjects, teacher effects on test scores are weak predictors of effects on non-test score outcomes (including GPA). This suggests that teacher test score effects measure certain skills, and teacher effects on the non-cognitive factor measure a largely *different* but potentially important set of skills. This indicates that many teachers who improve test scores may have average effects on non-test score outcomes, and many teachers who have large and important effects on

non-test score outcomes may have average effects on test scores. As indicated in the model, variability in outcomes associated with individual teachers that is unexplained by test scores may reflects unmeasured non-cognitive skills. If this is so, teacher effects on the non-cognitive factor might explain teachers' ability to improve long-run outcomes that are not measured by test scores. Section VI investigates this important possibility directly.

VI Predicting Longer Run Effects with Short Run Effects

While the relationships in Table 3 *suggest* that teachers who improve non-cognitive skills may also improve long-run outcomes, this section directly tests whether teachers who increase the non-cognitive factor actually *cause* students to have improved long-run outcomes (conditional on their test score effects). The main longer-run outcomes under study are measures of high school completion. To test this, I link predicted teacher effects to variables denoting whether the student subsequently dropped out of secondary school by 11th grade and graduated from high school by 12th grade (if one were able to observe completed schooling, one outcome would be 1 minus the other). I then test if students who have teachers that improve either test scores or the non-cognitive factor have better long-run outcomes. I estimate the equations below, where $\hat{\theta}_{j,test}$ and $\hat{\theta}_{j,noncog}$ are the predicted (out-of-sample) effects of teacher *j* on test scores and the non-cognitive factor, respectively. As before, θ_{sg} is a school track effect and θ_{sy} is a school-year effect. Standard errors are clustered at the teacher level.

$$Y_{ijcy} = A_{iy-1}\delta + \psi_{1,test}\hat{\theta}_{j,test} + X_{iy}\beta + I_{sgi}\,\theta_{sg} + \theta_{sy} + \varepsilon_{ijcy} \quad [9]$$

$$Y_{ijcy} = A_{iy-1}\delta + \psi_{2,test}\hat{\theta}_{j,test} + \psi_{2,noncog}\hat{\theta}_{j,noncog} + X_{iy}\beta + I_{sgi}\,\theta_{sg} + \theta_{sy} + \varepsilon_{ijcy} \quad [10]$$

To quantify the extent to which including both $\hat{\theta}_{j,noncog}$ and $\hat{\theta}_{j,test}$ in [10] increases our ability to predict variability in teacher effects over only including $\hat{\theta}_{j,test}$ in [9], I compute the percentage increase in the predicted variability of the teacher effects on the long-run outcome from [9] to [10]. Specifically, I compute $100\times\left(sd(\hat{\psi}_{2,test}\hat{\theta}_{j,test} + \hat{\psi}_{2,noncog}\hat{\theta}_{j,noncog}) / sd(\hat{\psi}_{1,test}\hat{\theta}_{j,test}) - 1\right)$. This is the percentage increase in how much of the teacher-level variability in the longer-run outcome can be detected by including teacher effects on the non-cognitive factor (over simply using test score value-added). The data for both subjects are stacked so that results are presented for all teachers. I present effects by subject in Section VI.2.

Column 1 of Table 7 shows that, on average, students with teachers who raise test scores by 1σ are 1.65 percentage points less likely to drop out of high school by 11th grade. This effect has the expected positive sign, is statistically significant at the 10 percent level, and is similar to the descriptive cross-sectional relationships presented in Table 3. To assuage concerns that the importance of the non-cognitive factor is driven by any one single non-test score outcome, columns 2 through 5 present the coefficient on the teacher effect on each non-test score outcome individually, all conditional on the test score effect. All of the effects on the non-test score outcomes have the expected signs: teachers who increase GPA reduce dropout; teachers who increase suspensions increase dropout; teachers who increase absences also increase dropout; and teachers who increase on-time grade progression decrease dropout. Teacher effects on GPA and on-time grade progression are both individually statistically significant at the 10 percent level — indicating that they each predict independent variation in dropout conditional on teachers test score effects. Combining all four non-test score outcomes into a single variable, column 6 shows that teacher effects on the non-cognitive factor have a statistically significant negative relationship with dropout. Specifically, students with a teacher who raises the non-cognitive factor by 1σ are 9.8 percentage points less likely to drop out of high school. Including the teacher effect on the non-cognitive factor increases the explained variability in teacher effects on dropout by 98 percent. To ensure that the results are robust to the full set of controls, column 7 includes school-by-year and school track effects. The estimates and conclusions are similar between columns 6 and 7.

To make sure that the estimated effects on dropout reflect real causal effects, I also test for selection on observables by estimating effects on predicted dropout (based on all observable student characteristics). The results in column 9 show that there is no relationship between predicted teacher effect and predicted dropout despite real effects on actual dropout. Testing for robustness to selection on unobservables, I regress the average school-track-cohort-level mean predicted teacher effect on dropout. The results are similar to those using individual-level data and indicate no selection on unobservables. The point estimates on test scores and the non-cognitive factor are surprisingly similar to those based on the student-level models that predict long run outcomes as a function of 9th grade outcomes with a rich set of controls. That is, having a teacher who raises test scores or the non-cognitive factor by 1σ has a similar effect to that of a 1σ difference in test scores or the non-cognitive factor across students within school (after conditioning on a rich set of student-level controls). This is reassuring because it is also one of the assumptions behind

the model presented in Section II.

Because the standard deviation of teacher effects on the non-cognitive factor is roughly 0.07, going from a teacher at the 15th to one at the 85th percentile of the non-cognitive effect distribution is associated with a 0.14*9.8=1.37 percentage point reduction in the likelihood of dropping out. In contrast, using the estimated test score effect by subject, going from a teacher at the 15th to one at the 85th percentile of the test score effect distribution is associated with a 0.07 percentage point reduction in the likelihood of dropping out for English teachers and a 0.17 percentage point reduction in the likelihood of dropping out for Algebra teachers. While these effects may seem modest, modest effects for a single student aggregated across all students in a class over their entire lifetime can result in important economic effects (Chetty et al., 2011).

The other measure of high school completion is high school graduation. Column 1 of Table 8 shows that, on average, students with teachers who raise test scores by 1σ are 3.97 percentage points more likely to graduate high school. This effect has the expected positive sign, is statistically significant at the five percent level, and is similar to the descriptive cross-sectional relationships presented in Table 3. Columns 2 through 5 present the coefficients on the teacher effects for each non-test score outcome individually conditional on test score effects. As with dropout, all of the effects on the non-test score outcomes have the expected signs. Moreover, teacher effects on GPA, being suspended, and on-time grade progression each individually predict independent variation in high school graduation conditional on teachers' test score effects. Looking to the combined factor, column 6 shows that teacher effects on the non-cognitive factor have a statistically significant positive effect on graduation. Specifically, students with a teacher who raises the non-cognitive factor by 1σ are 21.9 percentage points more likely to graduate from high school. Including teacher effects on the non-cognitive factor increases the explained variability of teacher effects on graduation by 86 percent. As with dropout, the results are robust to the full set of controls, are robust to aggregating the treatment to the school-track-cohort level, and there is no evidence of selection on observables. As with dropout, having a teacher who raises test scores or the non-cognitive factor by 1σ has a similar effect to that of a 1σ difference in test scores or the non-cognitive factor across students within schools (and conditional on a rich set of student controls). Computations show that going from a teacher at the 15th to one at the 85th percentile of the non-cognitive effect distribution is associated with a 0.14*21.9=3.06 percentage point increase in high school graduation. Also, going from a teacher at the 15th to one at the 85th percentile of the

test score effect distribution is associated with a 0.18 percentage point increase in graduating high school for English teachers and a 0.42 percentage point increase for Algebra teachers.

It is worth noting that if teachers have effects on dimensions of ability not captured by either their effects on test scores or the non-cognitive factor, these estimates may not capture a teacher's full effect on longer-run outcomes. However, it is clear that using both cognitive outcomes (e.g., test scores) and non-cognitive outcomes (e.g., the non-cognitive factor) increases our ability to identify excellent teachers who may improve longer run outcomes (rather than only increasing test scores).

VI.1 *Effects on Longer Run Outcomes by Subject*

The results thus far provide compelling evidence that (a) teachers have causal effects on soft skills (i.e., non-cognitive skills) that are associated with long term success and are not picked up by test scores, and (b) teachers who have real effects on non-cognitive skills also improve high school completion above and beyond what their effects on test score would predict. However, given that teachers have larger effects on Algebra test scores than English test scores, English teachers have larger effects on the non-cognitive factor than Algebra teachers, and English teacher effects on the non-cognitive factor are much more precisely estimated than those for Algebra teachers, it is reasonable to expect that the relative importance of teacher effects on the non-cognitive factor might be larger for English than for Algebra teachers. To test this, I estimate equations [9] and [10] separately for English and Algebra teachers. Results are presented in Table 9 both for models that include track school and year effects and also those that include track school effects and school-by-year effects.

Columns 1 and 2 present Algebra teacher effects on dropout, and columns 5 and 6 present Algebra teacher effects on high school graduation. All specifications indicate that Algebra teacher effects on test scores and the non-cognitive factor improve outcomes. Students who have 9th grade Algebra teachers that raise test scores by 1σ are between 1.6 and 1.8 percentage points less likely to drop out of high school and between 2.7 and 3.46 percentage points more likely to graduate high school. However, 9th grade Algebra teacher effects on the non-cognitive factor are not statistically significant in any model. This is not entirely surprising given that predicted Algebra teacher effects on the non-cognitive factor based on other years was only marginally statistically significant in predicting the non-cognitive factor itself. That is, the predicted Algebra teacher effects on the non-cognitive factor were relatively imprecise so that it is not surprising that the estimated effects on

dropout and graduation are also imprecise. Precision issues aside, including the Algebra teacher effect on the non-cognitive factor only increases the ability to predict Algebra teacher effects on dropout by about 20 percent and that for high school graduation by between 3 and 17 percent.

As expected, the results for English teachers suggest a much greater impact of teacher effects on the non-cognitive factor. Columns 3 and 4 show that students with 9th grade English teachers that raise scores by 1σ are between 1.3 and 2.2 percentage points less likely to drop out of high school. However, students with 9th grade English teachers that raise the non-cognitive factor by 1σ are about 9 percentage points less likely to drop out of high school. The relative importance of the non-cognitive factor is similarly pronounced for high school graduation. Columns 7 and 8 show that students with 9th grade teachers that raise English test scores by 1σ are about 2.7 percentage points more likely to graduate high school while students who have 9th grade English teachers that raise the non-cognitive factor by 1σ are 21.3 percentage points more likely to graduate high school. Looking at the ability to predict effects on the longer-run outcomes, including teacher effects on the non-test score outcomes increases the predictable variability of teacher effects on dropout by between 280 and 392 percent and increases the predictable variability of teacher effects on high school graduation by between 202 and 433 percent. While these increases may seem large, they are consistent with Chamberlain (2013) who finds that test score value-added accounts for less than one fifth of the overall effect of teachers on college-going. The effects by subject also provide an explanation for the Chetty et. al. (2011) finding that English teachers have larger effects on longer-run adult outcomes even though they have smaller effects on test scores than math teachers. If teacher effects on non-cognitive skills is more important for English teachers, this could explain their counterintuitive finding.

While high school dropout and graduation are the main long-run outcomes in this study, I also present effects of 9th grade teachers on the number of AP courses taken by 12th grade, whether a student took the SAT, and whether they expressed intentions to attend a four-year college in a high school graduation survey (Table 10). English teacher effects on the non-cognitive factor predict teacher effects on AP courses taken, SAT taking, and intentions to attend a four-year college, conditional on test score effects. These results are consistent with the finding that teacher effects on the non-cognitive factor improve our ability to identify teachers who improve longer-run outcomes considerably for English teachers, but little for Algebra teachers.

VI.2 *Are Teacher Effects on the Non-cognitive Factor and Test Scores Simply Different Measures of the Same Single Dimension of Ability?*

Given that teacher effects on test scores and teacher effects on the non-cognitive factor are positively correlated (albeit weakly), one may wonder if these are both measures of the same single dimension of ability. Specifically, if the value-added estimates reflect effects on students' unidimensional ability with error, then additional measures of the teacher effect on this same unidimensional ability will be positively correlated with the test score effect and may explain variability in the long-run effect unexplained by value-added.[17] Accordingly, it is important to know if the non-cognitive factor truly measures a different set of skills than test scores, or if test scores and the non-cognitive factor are noisy measures of the same set of skills. I present a test to tell these two scenarios apart. If the ability to predict effects on the long-run outcome were due to measurement error in the effect on test scores, then teacher effects on the non-cognitive factor should also increase our ability to predict effects on test scores conditional on a teacher's estimated test score effect (estimated out-of-sample). Intuitively, measurement error will lead one to understate both the relationship between test score effects and the effect on long-run outcomes *and* to understate the relationship between a teacher's test score effect (estimated out-of-sample) and her effect on test scores. As such, if measurement error is the explanation, then assuming that test scores and the non-cognitive factor measure the same dimensions of ability, we would expect teachers who improve the non-cognitive factor to improve students' test scores conditional on their out-of-sample test score effects.

To test this, I regress test scores on both out-of-sample predicted effects on the non-cognitive factor and out-of-sample predicted effects on test scores (with school track fixed effects and school-by-year fixed effects). Because the effects are driven by English teachers, I focus on the effect of English teachers on the non-cognitive factor and test scores. Conditional on test score effects, teacher effects on the non-cognitive factor yields a *negative* coefficient with a *p*-value 0.44 for English test scores. That is, teacher effects on the non-cognitive factor provide no additional predictive power for predicting English test scores (despite predicting much residual variation in teacher effects on longer-run outcomes). This is inconsistent with measurement error in value-

[17] From a policy perspective, what matters is that we can obtain a better prediction using the non-test score outcomes in conjunction with test scores. As such, it is irrelevant whether the additional predictive power of the effect on the non-cognitive factor is due to measurement error in the test score effects or due to test scores missing non-cognitive dimensions of ability. However, the distinction is economically meaningful.

added causing effects on the non-cognitive factor to explain effects on long-run outcomes. Instead, the results suggest that long-run effects reflect multiple dimensions of skills and that the non-cognitive factor captures dimensions of ability not well measured by test scores.

VI.2 *Correlations of Effects on the Non-Cognitive Factor and Possible Uses in Policy*

While the focus of this paper is the importance of accounting for effects on non-cognitive skills, in this section I briefly discuss practical uses for the non-cognitive factor in education policy. One policy use would be to identity those observable teacher characteristics associated with effects on the non-cognitive factor and select teachers with these characteristics. To determine the scope of this type of policy, I regress the non-cognitive factor on observable teacher characteristics (while controlling for school tracks, year effects, and student covariates). For Algebra teachers, observable teacher characteristics do not predict a large share of a teacher's effect on the non-cognitive factor. In fact, none of the primary characteristics — being fully certified, scoring well on teaching exams, having a regular license, and selectivity of a teacher's college — have a statistically significant relationship with the non-cognitive factor. Looking to experience, I include indicator variables for each year of teacher experience (from 0 to 29 years) and plot the experience profile for both the non-cognitive factor and Algebra test scores in the top panel of Figure 1. With more years of experience, test scores tend to improve, on average. The F-test of joint significance of all the teacher experience indicators yields a p-value of less than 0.001. However, for the non-cognitive factor the experience profile is much flatter. The F-test of joint significance of all teacher experience indicators yields a p-value of 0.62—suggesting no relationship between teacher experience and effects on the non-cognitive factor for Algebra teachers. Results for English teachers tell a similar story. The only observable teacher characteristic associated with improvements in the non-cognitive factor is scores on certification exams. Increasing a teacher's certification score by a standard deviation increases the non-cognitive factor by 0.0097σ. The experience profile in the lower panel of Figure 1 shows no statistically significant relationship between experience and effects on the non-cognitive factor. All in all, the observable teacher characteristics used in this research are not good predictors of teacher effects on non-cognitive skills measured by the factor. Accordingly, using observable teacher characteristics to identify excellent teachers may provide limited benefits.

Another policy application is to incentivize teachers to improve the non-cognitive factor. Because some of the outcomes that form the non-cognitive factor (such as grades and suspensions)

can be "improved" by changes in teacher behavior that do not improve student skills (such as inflating course grades, misreporting attendance, and leaving disciplinary infractions unreported) attaching external stakes to the non-cognitive factor may not improve students skills (even if the *measured* outcomes do improve). One possibility is to find measures of non-cognitive 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. As such, one could attach external incentives to both these measures of non-cognitive skills and test scores to promote better longer run outcomes.[18]

A final policy is to identify those teaching practices that cause improvements in the non-cognitive factor and encourage teachers to use these practices (through evaluation, training, or incentive pay). This avoids problems associated with "gaming" or rigging the outcomes by incentivizing observable, difficult-to-fake behaviors (such as asking questions or having group discussions) that may have causal effects on the non-cognitive factor. Such approaches have been used successfully in recent research to increase test scores (Taylor & Tyler, 2012). However, one could expand the model to identify best teacher practices based not only on test score gains but also gains in the non-cognitive factor. Indeed, the teacher evaluations systems designed by Allen et al. (2011) to promote teacher behaviors that lead to both improved test scores and better student-teacher interactions suggest that this may be a fruitful path.

VII Conclusions

This paper presents a two-factor model that assumes that all student outcomes are a function of both cognitive and non-cognitive ability. The model shows that one can use a variety of short-run outcomes to estimate a teacher's predicted effect on long-run outcomes, and that such outcomes would ideally reflect a combination of both cognitive and non-cognitive skills. In administrative data, a non-cognitive factor (a weighted average of non-test score student outcomes in 9th grade) is associated with sizable improvements in longer-run outcomes. Ninth grade English and Algebra teachers have meaningful effects on test scores, absences, suspensions, on-time 10th grade enrollment, and grades. Teacher effects on test scores and these non-test score outcomes (and the non-cognitive factor) are weakly correlated; many teachers who are among the best at

[18] A somewhat similar policy was suggested in the Gates Foundation report, Measures of Effective Teaching (MET). This multiple measure approach was proposed in Mihaly, McCaffrey, Staiger, & Lockwood (2013).

improving test scores may be among the worst at improving non-cognitive skills. Teacher effects on *both* test scores and the non-cognitive factor predict their effects on high school dropout rates, high school completion, SAT taking, and intentions to attend college. Indeed, teacher effects on the non-cognitive factor explain significant variability in their effects on these longer-run outcomes that are not captured by their test score effects. There are important difference across subjects such that adding teacher effects on the non-cognitive factor increases the predicted variability on longer-run outcomes by at least 200 percent for English teachers and only about 20 percent for Algebra teachers. While the specific measures of non-cognitive skills employed in this paper are by no means perfect (and there are likely much better measures that could be employed), the results highlight the broader point that using non-test score measures can be fruitful in evaluating human capital interventions.

The findings suggest that test score measures understate the effect of teachers on adult outcomes in general, and may greatly understate their importance in affecting outcomes that are determined by non-cognitive skills (such as dropping out, criminality, and being employed). The results provide hard evidence of an idea that many believe to be true but has never been shown concretely. That is, this study provides evidence that measuring teacher effects on test scores captures only a fraction of their effect on longer-run outcomes and presents the first evidence that evaluating teacher effects on non-test score outcomes may greatly improve our ability to predict teachers' overall effects on longer-run outcomes. More generally, this study highlights that a failure to account for the effect of educational interventions on non-cognitive skills can lead to biased estimates of the effect of such interventions on important longer-run outcomes. Finally, the analytic framework put forth in this paper can be used in other settings to estimate the effects of educational interventions through improvements in both cognitive and non-cognitive skills. Results from such analyses can then be used to identify practices that both increase test scores and improve non-cognitive skills.

Bibliography

1. Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*, 95-135.
2. Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to

enhancing secondary school instruction and student achievement. Science, 333, 1034-1037.
3. Alexander, K. L., Entwisle, D. R., & Thompson, M. S. (1987). School Performance, Status Relations, and the Structure of Sentiment: Bringing the Teacher Back In. *American Sociological Review, 52*, 665-82.
4. Altonji, Joseph G., Todd E. Elder & Christopher R. Taber, 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," Journal of Political Economy, University of Chicago Press, vol. 113(1), pages 151-184, February.
5. Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the Big Five in late childhood. *Personality and Individual Differences, 34*(4), 645-664.
6. Booker, K., Sass, T. R., Gill, B., & Zimmer, R. (2011). The Effect of Charter High Schools on Educational Attainment. *Journal of Labor Economics, 29*(2), 377-415.
7. Borghans, L., Weel, B. T., & Weinberg, B. A. (2008). Interpersonal Styles and Labor Market Outcomes. *Journal of Human Resources, 43*(4), 815-58.
8. Bowles, S., Gintis, H., & Osborne, M. (2001). The Determinants of Earnings: A Behavioral Approach. *Behavioral Approach, 39*(4), 1137-76.
9. Brookhart, S. M. (1993). Teachers' Grading Practices: Meaning and Values. *Journal of Educational Measurement, 30*(2), 123-142.
10. Carneiro, P., Crawford, C., & Goodman, A. (2007). The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes. *CEE Discussion Papers 0092*.
11. Cascio, E., & Staiger, D. (2012). Knowledge, Tests, and Fadeout in Educational Interventions. *NBER working Paper Number 18038*.
12. Chamberlain, Gary., "Predictive effects of teachers and schools on test scores, college attendance, and earnings" *PNAS 2013 ; October 7, 2013, doi:10.1073/pnas.1315746110*
13. Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. *Quarterly Journal of Economics, 126*(4), 1593-1660.
14. Chetty, R., Friedman, J., & Rockoff, J. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. (Unpublished manuscript).
15. Deming. (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*(3), 111-134.
16. Deming, D. (2011). Better Schools, Less Crime? *The Quarterly Journal of Economics, 126*(4), 2063-2115.
17. Downey, D., & Shana., P. (2004). When Race Matters: Teachers' Evaluations of Students' Classroom Behavior. *Sociology of Education, 77*, 267-82.
18. Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101.
19. Duckworth, A. L., & Carlson, S. M. (in press). Self-regulation and school success. In B.W. Sokol, F.M.E. Grouzet, & U. Müller (Eds.), *Self-regulation and autonomy: Social and developmental dimensions of human conduct*. New York: Cambridge University Press.
20. Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? : evidence from NELS88. *Industrial and Labor Relations Review, 48*, 547-561.
21. Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica, 77*(3), 953-973.
22. Fredriksson, P., Ockert, B., & Oosterbeek, H. (forthcoming). Long-Term Effects of Class Size. *Quartlerly Journal of Economics*.
23. Furnham, A., Monsen, J., & Ahmetoglu, G. (2009). Typical intellectual engagement, Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. *British Journal of Educational Phychology, 79*, 769-782.
24. Harris, D., & Anderson, A. (2012). Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence From Middle School Teachers. *Association of Public Policy Analysis & Management.* Baltimore.
25. Heckman, J. (1999). Policies to Foster Human Capital. *NBER Working Paper 7288*.
26. Heckman, J. J. and T. Kautz (2012, August). Hard evidence on soft skills. Lab. Econ. 19(4), 451–464. Adam Smith Lecture.
27. Heckman, J. J., & Rubinstein, Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review, 91*(2), 145-49.
28. Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics, 24*(3), 411-82.

29. Heckman, J., Pinto, R., & Savelyev, P. (forthcoming). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*.
30. Holmstrom, B., & Milgrom, P. (1991). Multitask Principal-Agent Analysis: Incentive Contracts, Asset Ownership and Job Design. *Journal of Law, Economics and Organization, 7*(Special Issue), 24-52.
31. Howley, A., Kusimo, P. S., & Parrott, L. (2000). Grading and the ethos of effort. *Learning Environments Research, 3*, 229-246.
32. Jackson, C. K. (2014). Teacher Quality at the High-School Level: The Importance of Accounting for Tracks. *Journal of Labor Economics*, Vol. 32, No. 4.
33. Jackson, C. K. (forthcoming). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics*.
34. Jackson, C. K., & Bruegmann, E. (2009). Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics, 1*(4), 85-108.
35. Jencks, C. (1979). *Who Gets Ahead? The Determinants of Economic Success in America.* New York: Basic Books.
36. Jennings, J. L., & DiPrete, T. A. (2010). Teacher Effects on Social and Behavioral Skills in Early Elementary School. *Sociology of Education, 83*(2), 135-159.
37. John, O., Caspi, A., Robins, R., Moffit, T., & Stouthamer-Loeber, M. (1994). The "Little Five": exploring the nomological network of the Five-Factor Model of personality in adolescent boys. *Child Development, 65*, 160–178.
38. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Pearson.
39. Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER working paper 14607*.
40. Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER Working Paper # 14607*.
41. Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-addedmodels. *Quantitative Economics, 3*, 333-362.
42. Koedel, C. (2008). An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Department of Economics, University of Missouri Working Paper 0808*.
43. Koedel, C. (2008). Teacher Quality and Dropout Outcomes in a Large, Urban School District. *Journal of Urban Economics, 64*(3), 560-572.
44. Koedel, C., & Betts, J. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? *Education Finance and Policy, 6*(1), 18-42.
45. Lindqvist, E., & Vestman, R. (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-128.
46. Lounsbury, J. W., Steel, R. P., Loveland, J. M., & Gibson, L. W. (2004). An Investigation of Personality Traits in Relation to Adolescent School Absenteeism. *Journal of Youth and Adolescence, 33*(5), 457–466.
47. Lucas, S. R., & Berends, M. (2002). Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking. *Sociology of Education, 75*(4), 328-348.
48. Mihaly, Kata., Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood (2013). "A Composite Estimator of Effective Teaching" Gates foundation Research Paper.
49. Mansfield, R. (2012). Teacher Quality and Student Inequality. (Working Paper) *Cornell University*.
50. Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73*(2), 417-458.
51. Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*.
52. Sadker, D. M., & Zittleman, K. (2006). *Teachers, Schools and Society: A Brief Introduction to Education.* McGraw-Hill.
53. Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance."*American Economic Review*, 102(7): 3628-51.
54. Waddell, G. (2006). Labor-Market Consequences of Poor Attitude and Low Self-Esteem in Youth. *Economic Inquiry, 44*(1), 69-97.

Tables and Figures

Table 1: *Summary Statistics of Student Data*

Variable	Mean	SD	SD within school-tracks	SD within schools
Math z-score 8th grade	0.251	(0.945)	(0.600)	(0.878)
Reading z-score 8th grade	0.237	(0.936)	(0.678)	(0.891)
Male	0.510	(0.500)	(0.482)	(0.498)
Black	0.259	(0.438)	(0.375)	(0.399)
Hispanic	0.068	(0.252)	(0.245)	(0.256)
White	0.579	(0.494)	(0.404)	(0.432)
Asian	0.020	(0.143)	(0.133)	(0.138)
Parental education: Some High School	0.065	(0.246)	(0.25)	(0.259)
Parental education: High School Grad	0.391	(0.487)	(0.454)	(0.474)
Parental education: Trade School Grad	0.017	(0.129)	(0.129)	(0.132)
Parental education: Community College Grad	0.136	(0.343)	(0.327)	(0.335)
Parental education: Four-year College Grad	0.235	(0.424)	(0.376)	(0.394)
Parental education: Graduate School Grad	0.072	(0.258)	(0.225)	(0.237)
Number of Honors classes	1.123	(1.405)	(0.575)	(1.163)
Algebra I z-Score (9th grade)	0.044	(0.988)	(0.775)	(0.889)
English I z-Score (9th grade)	0.049	(0.975)	(0.670)	(0.906)
Ln Absences	0.805	(1.106)	(0.927)	(0.984)
Suspended	0.048	(0.214)	(0.214)	(0.225)
GPA	2.913	(0.832)	(0.604)	(0.801)
In 10^{th} grade on time	0.898	(0.301)	(0.305)	(0.339)
Dropout (2005-2011 cohorts)	0.046	(0.211)	(0.205)	(0.213)
Graduate (2005-2010 cohorts)	0.825	(0.379)	(0.380)	(0.405)
Take SAT (2005-2009 cohorts)	0.404	(0.491)	(0.386)	(0.439)
Intend to attend college (2005-2010 cohorts)	0.400	(0.490)	(0.432)	(0.463)
Observations			464,502	

Notes: These summary statistics are based on students who took the English I exam. Incoming math scores and reading scores are standardized to be mean zero unit variance for al takers in that year. About 10 percent of students do not have parental education data—the missing category is "missing parental education". Readers might wonder why the estimation sample has average test scores in both 8^{th} and 9^{th} grade that are above the mean for all test takers in that grade in that year. For the 8^{th} grade scores, I take the last time a student was in 8^{th} grade. As such, the first attempt for students with low scores who also repeat 8^{th} grade are not included in the regression sample. Also, students with very low 8^{th} grade scores who repeat 8^{th} grade and never enroll in 9^{th} grade are not in the regression sample. Both these things lead to higher final 8^{th} grade scores among those who are in 9^{th} grade for the first time. The higher Algebra I and English scores for 9^{th} graders reflect the fact that those classrooms that could be matched to their teacher had slightly higher scores on average.

Table 2: *Correlations between the short run outcomes*

	Raw correlations between outcomes						Percentage of Variance Explained by Factors [a]		
	Log of # Days Absent	Suspended	Grade Point Average	In 10th grade on time	Algebra Score in 9th Grade	English Score in 9th Grade	Math Scores	English Scores	Non-cognitive Factor [b]
Ln of # Days Absent	1						0.025	0.036	0.468
Suspended	0.191	1					0.017	0.025	0.248
Grade Point Average	-0.276	-0.194	1				0.348	0.387	0.714
In 10th grade on time	-0.181	-0.151	0.447	1			0.089	0.095	0.400
Algebra Score in 9th Grade	-0.156	-0.128	0.590	0.294	1		1.000	0.381	0.243
English Score in 9th Grade	-0.097	-0.127	0.531	0.290	0.618	1	0.381	1.000	0.281

a. To obtain a measure of the percentage of the variance explains by test scores or the non-cognitive factor, I regress each short run outcome on test scores or the non-cognitive factor and report the R-squared

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

Table 3: *Relationship Between Short-run Outcome and Long-run Outcomes*

	1	2	3	4	5	6	7
	Dataset: NCERDC Micro Data						
	Main Longer Run Outcomes				Additional Outcomes		
	Drop out	Graduate	Drop out	Graduate	AP courses	Take SAT	Intend 4yr
Grade Point Average	-0.0431**	0.107**					
	[0.00107]	[0.00188]					
Log of # Absences	0.0116**	-0.0282**					
	[0.000437]	[0.000808]					
Suspended	0.0196**	-0.0440**					
	[0.00310]	[0.00514]					
On time in 10th grade	-0.0762**	0.295**					
	[0.00244]	[0.00416]					
English z-score	-0.00852**	0.0171**					
	[0.000788]	[0.00141]					
Math z-score	-0.00440**	0.00394*	-0.0123**	0.0238**	0.0592**	0.0968**	0.0575**
	[0.000927]	[0.00163]	[0.000768]	[0.00137]	[0.00172]	[0.00495]	[0.00166]
Non-cog factor z-score			-0.0571**	0.164**	0.168**	0.141**	0.139**
			[0.000867]	[0.00128]	[0.00128]	[0.00314]	[0.00122]
School Fixed Effects	Y	Y	Y	Y	Y	Y	Y
Covariates	Y	Y	Y	Y	Y	Y	Y
Observations	236,682	200,183	236,682	200,183	169,302	82,747	169,302

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

In addition to including school and year fixed effects, all models include controls for student gender, ethnicity, parental education, and a cubic function of Math and Reading test scores in 7th and 8th grade.

Table 4: *Effect of Out-of-Sample Estimated Teacher Effects and School-Track-Year-Level Mean Teacher Effects on Outcomes and Predicted Outcomes*

	1	2	3	4
	Algebra Teachers		English Teachers	
	Predicted Algebra Score	Predicted Non-cognitive Factor	Predicted English Scores	Predicted Non-cognitive Factor
Predicted Effect (*all variation*) [a]	0.00956	0.0204	0.00359	-0.0678
No student level controls	[0.0439]	[0.272]	[0.109]	[0.108]
	Algebra Scores	Non-cognitive Factor	English Scores	Non-cognitive Factor
Predicted Effect (*all variation*)	**1.000****	**1.000+**	**1.000****	**1.000****
	[0.0601]	**[0.602]**	**[0.0955]**	**[0.212]**
Predicted Effect (*within track cohorts*)	**0.981****	**1.024+**	**1.036****	**0.840****
	[0.0607]	**[0.589]**	**[0.104]**	**[0.203]**
Mean Predicted Effect (*across track cohorts*)[b]	**1.051****	0.911	**0.898****	**1.524****
	[0.124]	[1.073]	**[0.153]**	**[0.375]**
School Year Mean Effect (*across school tracks*)[b]	**0.964****	1.36	**1.085****	**1.337***
	[0.168]	[1.354]	**[0.245]**	**[0.634]**
Observations	193,677	193,677	316,322	316,322

Standard errors in brackets are adjusted for clustering at the teacher level. ** p<0.01, * p<0.05, + p<0.1

All models where teacher quality is defined at the individual teacher level include school-year effects and school-track fixed effects. The independent variable in within-cohort models is the predicted effect of a student's teacher (from all other years of data) on that outcome. The independent variable in the across-cohort models is the mean estimated effect (from all other years of data) of all students in the same school-track and the same cohort as the students for that outcome. The independent variable in the across-track models is the mean estimated effect (from all other years of data) of all students in the same school and the same cohort as the students for that outcome.

a. The predicted outcome reflects the effects of 7th and 8th grade test scores, parental education, gender, and ethnicity.
b. When the treatment is aggregated to the track or school level, the standard errors are clustered at the track and school level, respectively.

Table 5: *Estimated Covariance across Classrooms for the Same Teacher*

		Algebra Teachers						
		English Score	Algebra Score	Suspended	Log (1 + Absences)	GPA	In 10th on time	Non-cognitive Factor
Track-School FX and School Year FX	Implied SD	-	**0.0656**	0.0151	0.0445	**0.0368**	0.0099	**0.0725**
	95% Lower	0.0000	**0.0531**	0.0000	0.0000	**0.0097**	0.0000	**0.0513**
	95% Upper	0.0094	**0.0761**	0.0214	0.0762	**0.0511**	0.0211	**0.0888**
School FX and Year FX	Implied SD	0.0274	**0.1313**	**0.0375**	**0.1494**	**0.1154**	**0.0462**	**0.1691**
	95% Lower	0.0000	**0.1234**	**0.0330**	**0.1277**	**0.1065**	**0.0386**	**0.1568**
	95% Upper	0.0481	**0.1388**	**0.0416**	**0.1683**	**0.1236**	**0.0526**	**0.1805**

		English Teachers						
		English Score	Algebra Score	Suspended	Log (1 + Absences)	GPA	In 10th on time	Non-cognitive
Track-School FX and School Year FX	Implied SD	**0.0298**	-	**0.0152**	0.0392	**0.0418**	**0.0166**	**0.0648**
	95% Lower	**0.0155**	0.0000	**0.0109**	0.0000	**0.0328**	**0.0064**	**0.0515**
	95% Upper	**0.0392**	0.0283	**0.0186**	0.0614	**0.0492**	**0.0226**	**0.0758**
School FX and Year FX	SD	**0.0631**	**0.0895**	**0.0367**	**0.1554**	**0.0972**	**0.0407**	**0.1385**
	95% Lower	**0.0560**	**0.0811**	**0.0336**	**0.1444**	**0.0904**	**0.0363**	**0.1283**
	95% Upper	**0.0694**	**0.0972**	**0.0396**	**0.1657**	**0.1035**	**0.0446**	**0.1480**

Notes: The estimated covariances are computed by taking the classroom level residuals from equation 7 and computing the covariance of mean residuals 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 100 times and report the median estimated covariance as my sample covariance. To construct the standard deviation of this estimated covariance, I pair each classroom with a randomly chosen classroom under a different teacher and estimate the covariance. The standard deviation of 100 replications of these "placebo" covariances is my bootstrap estimate of the standard deviation of the estimated covariance. These two estimates are used to form confidence intervals for the covariance that can be used to compute estimates and confidence intervals for the standard deviation of the teacher effects (by taking the square root of the sample covariance and the estimated upper and lower bounds). In the two instances where the estimated covariance is negative, I report a missing value for the standard deviation. Note that under the null of zero covariance, one will have an estimated negative covariance half of the time. None of the negative covariances is statistically significantly different from zero at the ten percent level.

Table 6: *Proportion of the Variability in Estimated Effects Explained by Estimated Effects on Test Scores and Effects on the Non-cognitive Factor* [a]

	Algebra Test score effect	English Test score effect	Suspended Effect	Log of # Absences Effect	GPA Effect	On time enrollment in 10th grade Effect	Non-cognitive factor Effect
Algebra Test score FX	**1.00**	-	0.005	0.016	0.08	0.021	0.057
Non-cognitive factor FX	0.057	-	**0.348**	**0.500**	**0.635**	**0.379**	**1.00**
English Test score FX	-	**1.00**	0.001	0.008	0.025	0.016	0.027
Non-cognitive factor FX	-	0.027	**0.267**	**0.439**	**0.619**	**0.434**	**1.00**

a. This presents the estimated R-squared from separate regressions of a teacher's effect on each outcome on her effect on test scores and her effect on the non-cognitive factor. Estimates greater than 10 percent are in bold.

Table 7: *Predicting Effects on Dropout*

	1	2	3	4	5	6	7	8	9
Dependent Variable:	Dropout								Predicted Dropout
Level of Aggregation of Teacher Quality Measure:	Teacher level							Track cohort level	Teacher level
	Main regression Models							Specification Checks	
Effect: Test Score	**-0.0165+**	-0.0108	-0.0159	**-0.0160+**	-0.012	-0.0126	-0.0142	-0.0183	-0.00096
	[0.00984]	[0.00933]	[0.00982]	**[0.00964]**	[0.00972]	[0.00952]	[0.00921]	[0.0204]	[0.00128]
Effect: GPA		**-0.0216+**							
		[0.0111]							
Effect: Suspended			0.0237						
			[0.0244]						
Effect: Absences				0.00252					
				[0.00630]					
Effect: In 10th on time					**-0.0686****				
					[0.0173]				
Effect: Non-cognitive						**-0.0982***	**-0.0829+**	**-0.1787***	0.00304
						[0.0432]	**[0.0442]**	**[0.0743]**	[0.0055]
School-Track Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
School Year Effects	N	N	N	N	N	N	Y	N	N
Year Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations						442,823			
% increase in variance						97.99%	69.25%		

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1

Table 8: *Predicting Effects on High School Graduation*

	1	2	3	4	5	6	7	8	9
Dependent Variable:	Graduate								Predicted Graduate
Level of Aggregation of Teacher Quality Measure:	Teacher level							Track cohort level	Teacher level
	Main regression Models							Specification Checks	
Effect: Test Score	**0.0397*** **[0.0180]**	0.0287 [0.0183]	**0.0373*** **[0.0180]**	**0.0371*** **[0.0180]**	**0.0339+** **[0.0180]**	**0.0307+** **[0.0180]**	**0.0293+** **[0.0176]**	0.0465 [0.0368]	0.0035 [0.0034]
Effect: GPA		**0.0403+** **[0.0220]**							
Effect: Suspended			**-0.0948*** **[0.0458]**						
Effect: Absences				-0.0137 [0.0121]					
Effect: In 10th on time					**0.0831*** **[0.0348]**				
Effect: Non-cognitive						**0.219**** **[0.0732]**	**0.188**** **[0.0720]**	**0.319*** **[0.162]**	-0.0151 [0.0154]
School-Track Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
School Year Effects	N	N	N	N	N	N	Y	N	N
Year Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	369,590								
% increase in variance explained						86.43%	75.49%		

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1

Table 9: *Teacher Effects on High School Completion by Subject*

	1	2	3	4	5	6	7	8
	Algebra		English		Algebra		English	
	Dependent Variable: Dropout				Dependent Variable: Graduate			
Effect: Non-cognitive	-0.117	-0.134	**-0.0996***	**-0.0893+**	0.1960	0.0772	**0.213****	**0.213****
	[0.114]	[0.118]	**[0.0493]**	**[0.0524]**	[0.268]	[0.246]	**[0.0747]**	**[0.0736]**
Effect: Test Score	**-0.0166+**	**-0.0180+**	0.0130	0.0221	0.0276	**0.0346+**	0.0266	0.00279
	[0.0108]	**[0.0104]**	[0.0325]	[0.0288]	[0.0203]	**[0.0203]**	[0.0486]	[0.0470]
Track by School FX	Y	Y	Y	Y	Y	Y	Y	Y
Year FX	Y	Y	Y	Y	Y	Y	Y	Y
School-by-Year FX	N	Y	N	Y	N	Y	N	Y
Observations	164,546	164,546	278,277	278,277	135,398	135,398	234,192	234,192
% Increase	21%	19%	2805%	392%	17%	3%	202%	433%

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1

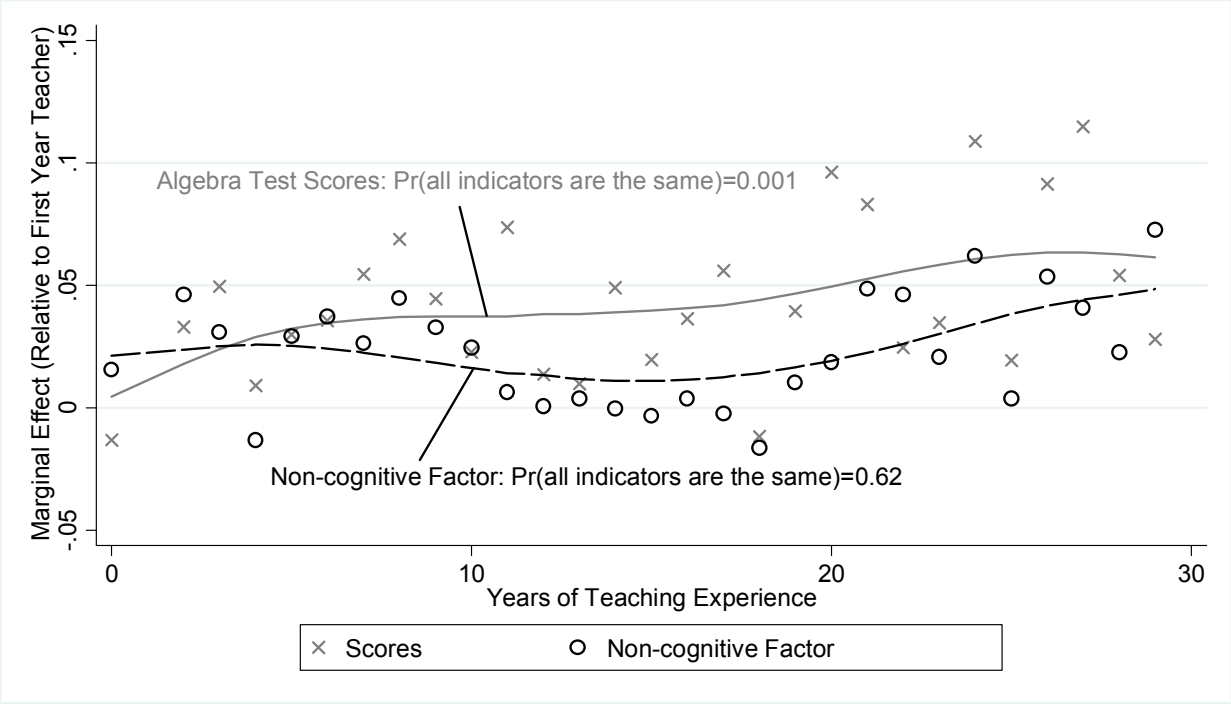
Table 10: *Teacher Effects on Predictors of College-Going by Subject*

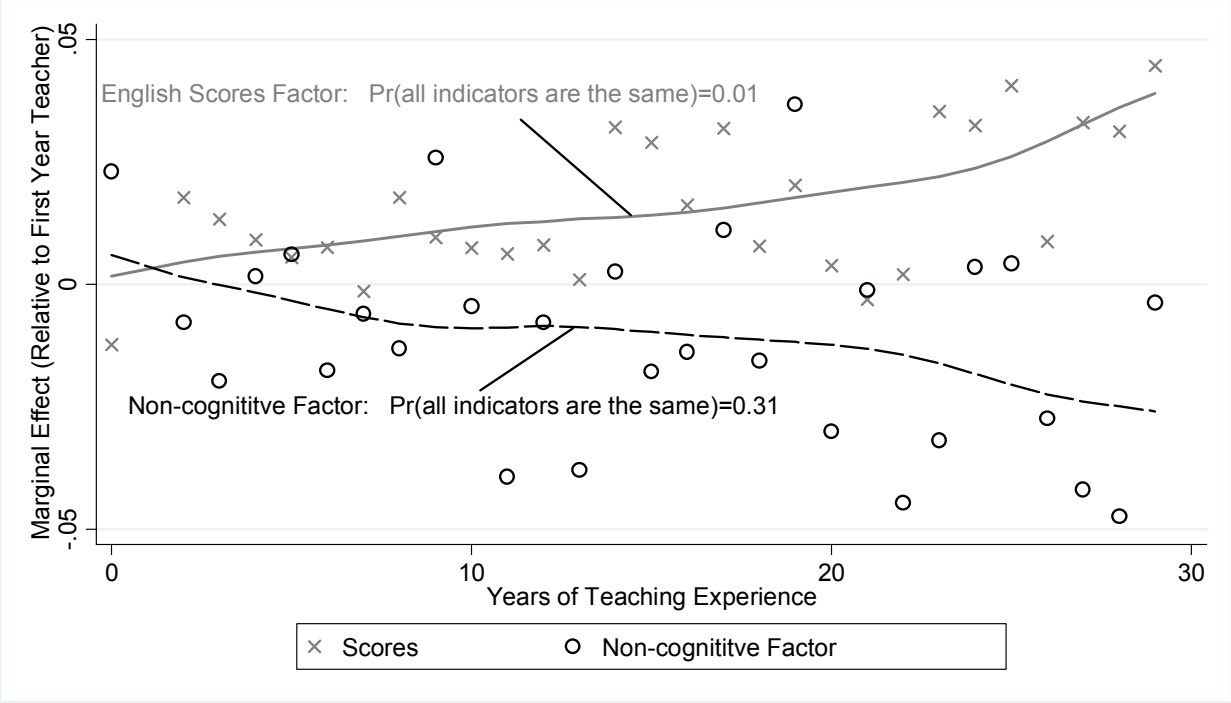
	1	2	3	4	5	6
	Algebra	English	Algebra	English	Algebra	English
	AP Courses Taken		SAT Taker		Intend 4-Year College	
Effect: Non-cognitive	-0.322	**0.664+**	-0.11	**0.112+**	0.33	**0.137+**
	[0.646]	**[0.354]**	[0.239]	**[0.0683]**	[0.236]	**[0.0777]**
Effect: Test Score	0.00651	0.0277	0.0189	-0.00571	**0.0308+**	-0.00618
	[0.0612]	[0.215]	[0.0278]	[0.0762]	**[0.0179]**	[0.0544]
Track by School FX	Y	Y	Y	Y	Y	Y
Year FX	Y	Y	Y	Y	Y	Y
Observations	61,645	95,443	135,398	234,192	135,398	234,192
% Increase in Explained Variance	11%	60%	21%	813%	81%	821%

Robust standard errors in brackets are adjusted for clustering at the teacher level.
** p<0.01, * p<0.05, + p<0.1

Figures

Figure 1: *Effect of Experience on Test Scores and Non-cognitive Factor*





Appendix

Appendix Note 1: *Matching Teachers to Students*

The teacher ID in the testing file corresponds to the teacher who administered the exam, who is not always the teacher that taught the class (although in many cases it will be). To obtain high-quality student-teacher links, I link classrooms in the End of Course (EOC) testing data with classrooms in the Student Activity Report (SAR) files (in which teacher links are correct). The NCERDC data contains End of Course (EOC) files with test-score-level observations for a certain subject in a certain year. Each observation contains various student characteristics, including ethnicity, gender, and grade level. It also contains the class period, course type, subject code, test date, school code, and a teacher ID code. Following Mansfield (2012), I group students into classrooms based on the unique combination of class period, course type, subject code, test date, school code, and the teacher ID code. I then compute classroom-level totals for student characteristics (class size, grade level totals, and race-by-gender cell totals). The Student Activity Report (SAR) files contain classroom-level observations for each year. Each observation contains a teacher ID code (the actual teacher in the course), school code, subject code, academic level, and section number. It also contains the class size, the number of students in each grade level in the classroom, and the number of students in each race-gender cell.

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

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

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

Appendix Note 2: *Estimating Efficient Teacher Fixed Effects*

I follow the procedure outlined in Kane and Staiger (2008) to compute efficient teacher fixed effects. This approach accounts for two issues: (1) teachers with larger classes tend to have more precise estimates and (2) there are classroom-level disturbances so that teachers with multiple classrooms will have more precise estimates. As before, I compute mean residuals from [7] for each classroom $\overline{e}^{*}_{cj} \equiv \theta_j + \phi_c + \hat{\varepsilon}_c$. Since the classroom error is randomly distributed, I use the covariance between the mean residuals of classrooms for the same teacher $\mathrm{cov}(\overline{e}^{*}_{cj}, \overline{e}^{*}_{c'j}) = \hat{\sigma}^2_{\theta_j}$ as an estimate of the variance of true teacher quality. I use the variance of the classroom demeaned residuals as an estimate of $\hat{\sigma}^2_{\varepsilon}$. Because the variance of the residuals is equal to the sum of the variances of the true teacher effects, the classroom effects, and the student errors, I compute the variance of the classroom errors σ^2_c by subtracting σ^2_{ε} and $\hat{\sigma}^2_{\theta_j}$ from the total variance of the residuals. For each teacher I compute [A1], a weighted average of their mean classroom residuals, where classrooms with more students are more heavily weighted in proportion to their reliability.

$$\hat{\theta}_j = \sum_{t=1}^{T_j} z_{jt} \cdot \frac{(1/(\sigma^2_c + (\sigma^2_{\varepsilon}/N_c))}{\sum_{t=1}^{T_j}(1/(\sigma^2_c + (\sigma^2_{\varepsilon}/N_c))} \qquad \text{[A1]}$$

Where N_c is the number of students in classroom *c*, and T_j is the total number of classrooms for teacher *j*. This is a more efficient estimate of the teacher fixed effect that the simple teacher average.

Appendix Note 3: *Analysis of the NELS-88 data*

To ensure that the patterns are not specific to North Carolina, I also employ data from the National Educational Longitudinal Survey of 1988 (NELS-88). The NELS-88 is a nationally representative sample of respondents who were eighth-graders in 1988. Table A3 presents the same models using the NELS-88 data. The results are largely consistent with those from the NCERDC data. For both dropout and high school graduation, the marginal effect of a 1σ increase in the non-cognitive factor is associated with marginal effects that are more than 10 times larger than that associated with a 1σ increase in math scores. Also similar to the NCERDC data, the results for college-going show much more similar predictive ability for test scores and the non-cognitive factor. A 1σ increase in test scores is associated with a 4.5 percentage point increase in college-going while a 1σ increase in the non-cognitive factor is associated with a 9 percentage point increase (an effect twice that of test scores).

The NELS-88 data also include longer-run outcomes from when the respondent was 25 years old. These allow one to see how this non-cognitive factor (based on 8th grade outcomes) predicts being arrested (or having a close friend who was arrested), employment, and labor market earnings, conditional on 8th grade test scores. The results show that test scores do not predict being arrested, but a 1σ increase in the non-cognitive factor is associated with a 4.5 percentage point decrease in being arrested (or having a close friend who was arrested). In contrast, both test scores and the non-cognitive factor predict employment in the labor market and earnings. Specifically, a 1σ increase in test scores is associated with a 1.18 percentage point increase in working, while a 1σ increase in the non-cognitive factor is associated with a similar 1.53 percentage point increase. Finally, conditional on having any earnings, a 1σ increase in test scores is associated with 13.8 percent higher earnings while a 1σ increase in the non-cognitive factor is associated with 20 percent higher earnings.

In recent findings, both Lindqvist & Vestman (2011) and Heckman, Stixrud, & Urzua (2006) find that non-cognitive ability is particularly important at the lower end of the earnings distribution. Insofar as the non-cognitive factor truly captures non-cognitive skills, one would expect this to be the case for this factor also. To test for this, I estimate quantile regressions to obtain the marginal effect on log wages at different points in the earnings distribution. The results (appendix table A4) show that at the 90th percentile through the 75th percentile of the earnings distribution, a 1σ increase in test scores and the non-cognitive factor is associated with a very similar increase of about 6 percent higher earnings. However, at the median level the non-cognitive factor is more important; the marginal effect of a 1σ increase in test scores and the non-cognitive factor is 3.8 percent and 9 percent higher earnings, respectively. At the 25th percentile, this difference is even more pronounced. A 1σ increase in test scores is associated with 2.6 percent higher earnings while a 1σ increase in the non-cognitive factor is associated with 17 percent higher earnings. These findings are similar to those by Lindqvist & Vestman (2011), suggesting that this factor is a reasonable measure of non-cognitive ability.

Table A1: *Most common academic courses*

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

Table A2: *Distribution of Number of Teachers in Each School-Track-Year Cell*

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

Note: This is after removing singleton tracks.

Table A3: *Relationship Between Short-run Outcome and Longer-run Outcomes*

	1	2	3	4	5	6
	Dataset: National Educational Longitudinal Survey 1988					
	Dropout	Graduate	College	Arrests	Working	Log Income
Math z-score	0.00326	0.00334	**0.0454****	0.0112+	**0.0118***	**0.138****
	[0.00242]	[0.00399]	**[0.00536]**	[0.00582]	**[0.00484]**	**[0.0486]**
Non-cog factor z-score	**-0.0222****	**0.0776****	**0.0905****	**-0.0454****	**0.0153****	**0.200****
	[0.00238]	**[0.00397]**	**[0.00479]**	**[0.00515]**	**[0.00434]**	**[0.0433]**
School Fixed Effects	Y	Y	Y	Y	Y	Y
Covariates	Y	Y	Y	Y	Y	Y
Observations	10,792	10,792	10,792	10,792	10,792	10,792

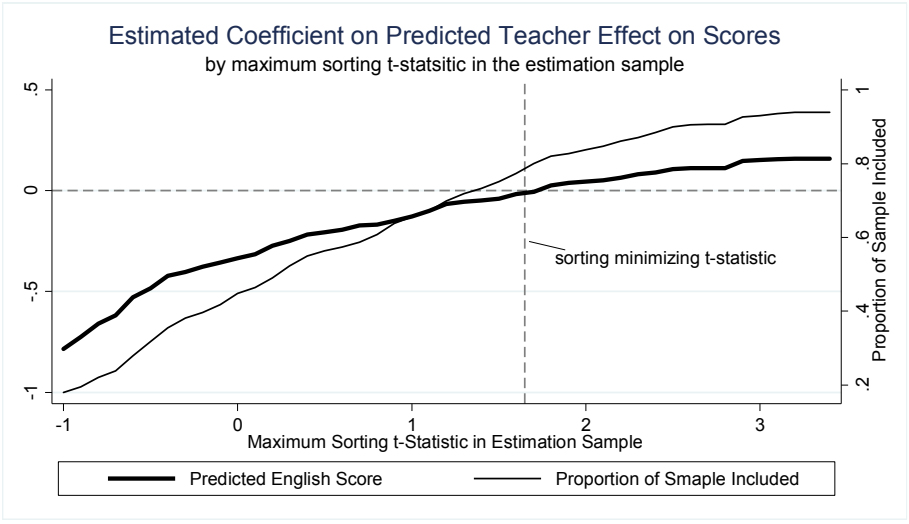
Robust standard errors in brackets
** p<0.01, * p<0.05, + p<0.1

Table A4: *Effect of test scores and the non-cognitive factor in 8th grade on adult earnings at different percentiles (NELS-88 sample)*

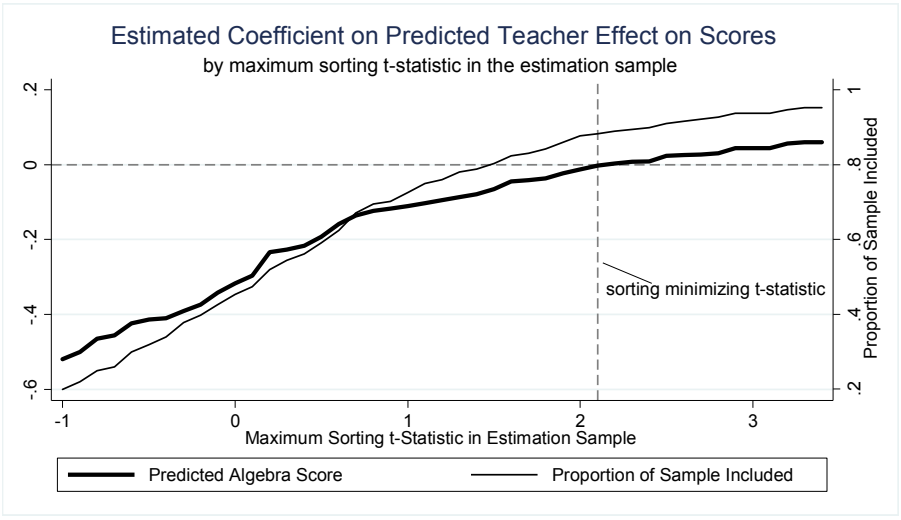
	Natural log of Income			
Percentile	25th	50th	75th	90th
Math z-score	0.0264	0.0382***	0.0512***	0.0562***
	[0.0481]	[0.00906]	[0.00667]	[0.00877]
Non-cog factor	0.174***	0.0906***	0.0705***	0.0619***
	[0.0462]	[0.00870]	[0.00641]	[0.00843]
Observations	10,792	10,792	10,792	10,792

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Appendix Figure A1



a.



b.

Appendix Note 4: *Validity of the sample restriction method for selection on observables*

One might wonder whether removing schools that exhibit the strongest positive selection of student ability to teacher quality until there is no correlation between teacher quality and observed student ability in the remaining sample yields unbiased estimates of the effect of teacher quality on outcomes. To see if this procedure yields an unbiased estimate of the teacher effect on student outcomes (under the assumption of selection on observables), I ran a Monte Carlo simulation of this procedure on simulated data.

I created 600 schools, each observed for six years, and each with the same 4 teachers across all years. Each classroom (teacher year) has 35 students, resulting in 504,000 student observations. Under the simulation, each school has a random fixed effect θ_s and a random selection slope δ_s. The random school-level selection slope determines the correlation between student ability and teacher quality within each school. As such, some schools have positive selection of teacher quality to student ability while others have negative selection or no selection. Each teacher has a random fixed effect θ_j, and each year has a random fixed effect τ_t. The random student-level error, ε_{ijs}, is the average of a selection term, $(\delta_s+\pi)\times\theta_j$, which is the random school-level slope (plus a constant) times the teacher effect, and an idiosyncratic student specific error term θ_i, and is defined as [1].

[1] $\varepsilon_{ijs} = ((\delta_s+\pi)\times \theta_j + \theta_i)/2$

The constant π is greater than zero such that there are more schools with positive selection than those with negative selection (as is the case in the data). The value chosen was 0.5. For simplicity θ_s, δ_s, θ_j, τ_t and θ_j are all drawn from a normal distribution with zero mean and unit variance.

The overall student outcome is the school effect plus the year effect plus the teacher effect plus the student-level error (which includes both a selection piece and an idiosyncratic piece).

[2] $Y_{ijst} = \theta_s + \theta_j + \tau_t + \varepsilon_{ijs}$

On this simulated data, I regress the student error term, ε_{ijs}, on the teacher fixed effect, θ_j, for each school and then compute the t-statistic of the slope for each school (note: this is analogous to regressing the predicted outcome on the predicted teacher effect and taking the t-statistic of the slope). I then remove the schools with the largest t-statistics until the coefficient on the teacher effect in predicting the student term is effectively zero (i.e. within some epsilon band of zero). Using the remaining "restricted" sample, I regress the outcome, Y_{ijst}, on the teacher effect and report the coefficient. If this procedure of dropping schools with strong positive sorting is valid, the coefficient on the teacher effect should be close to 1 in the restricted sample of schools even if there is sizable bias in the full sample of all schools. To asses this, I ran the simulation and estimated the coefficient 100 times and then plotted the distribution of the estimated coefficients across these 100 replications in Figure A2.

The top panel of Figure A2 plots the distribution of the coefficient on the teacher effect for the full sample under sorting. The lower panel plots both (a) the distribution of the coefficient on the teacher effect for the full sample without any sorting, and (b) the distribution of the coefficient on the teacher effect using the restricted sample under sorting.

Using the full sample under sorting, the estimated coefficients on teacher quality are centered on 1.233, and none of the 100 replications yields an estimate close to 1. In contrast, in the restricted sample under sorting, the coefficients are all close to 1 and are centered on 1. For comparison purposes the distribution of estimates under no sorting is also presented. While the spread of the coefficients around 1 is larger for the restricted sample than one would obtain using the full sample under no sorting, the restricted sample procedure clearly results in unbiased teacher

quality effect estimates under selection on observables. While the Monte Carlo simulation indicates that this procedure addresses the problem of selection on *obervables*, this procedure may not address the problem of selection on *unobservables*. To address the problem of selection on unobservables, I implement additional tests in Section IV.2.b.

Figure A2: *Distribution of Teacher Effect Estimates on Simulated Data using the Full Sample and the Restricted Sample under School-Varying Selection of Students to Teachers*