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

C. Kirabo Jackson, August 2, 2013
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). Teachers have causal effects on certain non-cognitive skills not measured by testing, but reflected in absences, suspensions, grades, and on-time grade progression. Measuring teacher effects on a weighted average of these non-test score outcomes (a proxy for non-cognitive skills) predicts effects on dropout, SAT-taking, and college plans—above and beyond their effects on test scores. Accordingly, test scores alone fail to identify many excellent teachers and may understate the long-run importance of teachers. (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 development), 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*; "a more comprehensive evaluation 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 also agree that teachers are an important component of the schooling environment. Studies show that having a teacher who ranks in the 85th percentile 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 &

[1] I thank David Figlio, Jon Guryan, Simone Ispa-Landa, Clement Jackson, Mike Lovenheim, James Pustejovsky, and Steven Rivkin for insightful and helpful 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.

Staiger, 2008; Rivkin, Hanushek, & Kain, 2005). The focus on test scores is largely because they are typically the best available measure. However, the research on non-cognitive skills provides reason to suspect that effects on test scores may in fact 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 non-test score outcomes. Finally, it tests whether 9th grade teacher effects on a proxy for non-cognitive ability predict effects on longer-run outcomes (e.g. high school completion and college exam-taking) conditional on test score effects. The resulting estimates and data are used to gauge 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 the first to document that teacher effects on non-cognitive outcomes (unrelated to value-added) predict important long-run effects.[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 measure only one dimension of ability. If teachers improve skills not captured by test-scores, then (a) many excellent teachers who improve long-run outcomes may not raise test scores, (b) the ability to raise test scores may not be the best predictor of effects on long-run outcomes, and (c) a regime that emphasizes test scores might induce teachers to divert effort away from skills not

[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. However, these studies may reflect differences in teacher perception rather than actual student behavior. 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.

captured by test scores to increase test score outcomes – potentially decreasing teacher quality overall (Holmstrom & Milgrom, 1991). This paper speaks to the second critique by assessing whether teachers affect skills not captured by test scores, and whether teacher effects on a proxy for non-cognitive skills predict long-run outcomes (conditional on their 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 (all in 9th grade) predict high school dropout, graduation, and SAT taking, 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 predictions of the model and investigates the extent to which teacher effects on non-cognitive outcomes predict effects on high school dropout, high school completion, SAT taking, and college aspirations above and beyond that predicted by their test score effects alone.

The results show that much of the variability in absences, suspensions, grades, and grade progression is uncorrelated with test scores. Consistent with this, an underlying non-cognitive factor that explains covariance across these non-test score outcomes (i.e. a weighted average of these non-test score outcomes) has a moderate correlation with test scores. This non-cognitive factor is associated with less high school dropout, increased high school graduation, and more SAT test-taking (a good proxy for college attendance), 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 non-cognitive ability not well measured by test scores.

Based on administrative data, 9th grade algebra and English teachers have meaningful effects on test-scores, non-test score outcomes, and the estimated non-cognitive factor. To address problems associated with student tracking in secondary school, this paper follows Jackson (forthcoming) and estimates models that condition on a student's school-track (the unique combination of school and specific courses taken) so that comparisons are made among students at the same school and in the same academic track, thus precluding bias due to student selection to tracks or treatments that vary across tracks. To show the validity of these models, the quasi-experimental tests proposed in Chetty et al. (2011) are employed, and they indicate little to no selection bias. Results indicate that teacher effects on test scores and the non-cognitive factor have a weak positive correlation, so that many teachers that increase the non-cognitive factor do not raise test scores and *vice versa*. Teacher effects on test scores predict effects on school dropout, high school completion, SAT taking, and college plans at graduation. However, teacher effects on the non-cognitive factor also predict effects on longer run outcomes conditional on their test score effects. Including teacher effects on the non-cognitive factor increases the estimated variability of teacher effects on these longer run outcomes by between 30 and 700 percent.

These findings are the first to demonstrate that non-cognitive outcomes can identify teachers who improve longer-run outcomes but are no more effective than average at raising test scores. While test score value-added is an important tool in identifying quality teachers, these results support a broader and more holistic view of student well-being and teacher quality. Also, evidence that teachers have effects on abilities not measured by test scores offers a potential explanation for interventions with test score effects that “fade out” over time that have lasting effects on adult outcomes (Chetty et. al. 2011; Cascio & Staiger, 2012).

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. Section V analyzes short run teacher effects. Section VI analyzes how short run teacher effects predict longer-run teacher effects, and 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 $E[\omega] = (0,0)$, that 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*. Figure 1 presents the path diagram.

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 may be correlated with effects on multiple short-run outcomes, even if effects on these short-run outcomes are not correlated 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.

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

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

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

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 where ability is multidimensional, there may be improvements in our ability to predict teacher effects on long-run outcomes by evaluating teacher effects on multiple outcomes that reflect a variety of skills (over a single outcome).[5] Intuitively, with uni-dimensional 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 reflects dimensions of ability not captured by the first outcome. If the second outcome reflects

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

[5] Note that this could also be true if short run outcomes were measured with error in a unidimensional model. That is, if outcomes 1 and 2 both measure the same dimension of ability with random noise, then 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 V, I demonstrate that this is unlikely to be the case for the outcomes used in this paper.

different abilities from the first, with multidimensional ability, 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 2010 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 the endogeneity 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 348,547 students in 619 secondary schools, 4296 English I teachers, and 3527 Algebra I teachers. The gender split is roughly even. About 58 percent of the 9th graders are white, 29 percent are black, 7.5 percent are Hispanic, 2 percent are Asian, and the remaining one percent are Native American, mixed race, or other. Regarding the parental highest education level (i.e. the highest level of education obtained by either of the student's two parents), about 7.5 percent were below high-school, 40 percent had a high school degree, about 15 percent had a junior college or trade school degree, 20 percent had a four-year college degree or greater, and 6.4 percent had an advanced degree (about 10 percent of students are missing data on parental education.) The test score variables have been standardized to be mean zero with unit variance for each cohort and test. Incoming 7th and 8th grade test scores of students in the sample are about 8 percent of a standard deviation higher than that of the average in 7th or 8th grade. This is because the sample of first time 9th grade students is less likely to have repeated a grade or to have dropped out of the schooling system. Data on high school dropout, high school graduation, SAT taking, and self-reported intentions to attend college upon graduation (available for years 2008 through 2012), are

[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.

linked to the 9th grade cohorts for years 2005 through 2009. In the sample, roughly 8.3 percent of 9th graders dropped out of school, while 79.3 percent graduated from high school. The remaining 11 percent either transferred out of the North Carolina school system or could not be tracked. Among 9th graders, roughly 39.3 took the SAT by 12th grade and 38.7 report having intentions to attend college after graduation.

Correlations among the short run outcomes

The correlations between the 9th grade outcomes reveal some interesting patterns. The first pattern is that test scores are relatively strongly correlated both with each other (math scores and reading scores have correlation≈0.6) and with grade point average (correlation≈0.55), but are weakly correlated with other non-test-score outcomes. Specifically, the correlations between the natural log of absences is -0.098 for algebra test scores and -0.082 for English test scores, and the correlations between being suspended is -0.13 for both algebra and English test scores. While slightly higher, the correlation between on-time progression to 10th grade and test scores is only 0.31. 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. Indeed, Table 2 indicates that test scores predict less than 2 percent of the variability in absences and suspensions, under 10 percent of the variability in on-time grade progression, and only about one third of the variability in GPA. Insofar as these are outcomes of interest in their own right, test scores may not measure *overall* educational or academic well-being.

The second notable pattern is that *most* behavioral outcomes are more highly correlated with each other than with test scores. Specifically, the correlation between suspensions and test scores is half that between suspensions and absences. Also the correlation between absences and test scores is about half that of the correlations between on time grade progression, GPA, or being suspended. The third notable pattern is that GPA is relatively well correlated with both the test score and the more behavioral 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 grades reflect a combination of cognitive and non-cognitive skills. In sum, in the context of the model, the patterns indicate that there are three groups of variables: those that are mostly cognitive (English and algebra 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) both higher test scores and this estimated non-cognitive factor are independently associated with better adult outcomes.

Table 3 shows that both test scores and non-test score outcomes independently predict long-run outcomes. I regress long-run outcomes (in 11th and 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 through 3 show that while higher test scores in 9th grade do predict less dropout, more high school graduation, and increased SAT taking, 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, more high school graduates, and more SAT taking. Similarly, increased suspensions and absences are associated with increased dropouts, lower high school graduation, and less SAT taking. For all three 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 1 percent level.

Because it is difficult to interpret multiple non-test-score outcomes in the same model, I created a weighted average of the non-test score outcomes as a single proxy for non-cognitive ability. To do this, I estimated a factor model on the four non-test-score outcomes (GPA, absences, suspensions, and on-time grade progression) and computed the unbiased prediction of the first underlying factor as my proxy for non-cognitive ability.[7] This average was then standardized to be mean zero unit variance. This weighted average is an estimate of the underlying ability that

[7] This predicted factor is, Factor = -0.35*suspended – 0.4*absenses+0.5*on time in 10th grade + 0.6*GPA.

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 28 percent of the variability in absences, 36 percent of the variability in being suspended, 67 percent of the variability in GPA, and 56 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 23 and 27 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 4, 5, and 6 show that for all three longer-run outcomes, a standard deviation increase in the non-cognitive factor is associated with larger improvements than a standard deviation (σ) increase in test scores (results are similar and slightly smaller using English test scores.) Specifically, while a 1σ increase in test scores is associated with a 0.7 percentage point decrease in dropout, a 1σ increase in the non-cognitive factor is associated with a 4.85 percentage point decrease in dropout. Also, while a 1σ increase in test scores is associated with a 1.8 percentage point increase in high school graduation, a 1σ increase in the non-cognitive factor is associated with an 18.2 percentage point increase. The predictive ability for test scores and the non-cognitive factor in terms of SAT taking is more similar; a 1σ increase in the non-cognitive factor is associated with a 15.2 percentage point increase in SAT taking, while a 1σ increase in test scores is associated with a somewhat smaller 8.89 percentage point increase in SAT taking. These numbers suggest that non-cognitive ability (as proxied by the factor) is a better predictor of dropout and high school graduation than test scores, and an equally good predictor for SAT taking.

To validate the use of the factor, I replicate the results in Table 3 using data from 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 8th grade. Looking to other adult outcomes, 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 their test scores. As found in other studies, the non-cognitive factor explains more variability in adult outcomes than test scores. 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 higher 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, a higher likelihood of 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 abilities 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 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 important 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 teacher effects on student outcomes. The empirical approach taken is to model student outcomes as a function of lagged student achievement and other student covariates, with the additional inclusion of controls for student selection to tracks and any treatments specific to tracks that might affect student outcomes directly. I do this by including indicators for a student's academic track.[8] Following Jackson (forthcoming), I define a school track as the unique combination of the 10 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] Defining tracks flexibly at the school/course-group/course level allows for different schools that have different selection models and treatments for each track.

The key identifying assumption is that students are randomly assigned to teachers within tracks. 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. To accomplish this, I model the outcomes Y_{icjgys} of student *i* in class *c* with teacher *j* in school track *sg*, at school *s*, in year *y* with [4] below (note: most teachers are observed in multiple classes).

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

A_{iy-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_{sgi} is an indicator variable equal to 1 if student *i* is in school track s*g* and *0* otherwise so that θ_{sg} is a time-invariant fixed effect for school track s*g*; 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_{ijgy} is a mean zero random error term.

By conditioning on school-tracks, one can obtain consistent estimates of the teacher effects θ_j as long as there is no selection to teachers *within* a school track. In these models, the teacher effects are teacher-level means of the outcome after adjusting for incoming student characteristics,

[8] 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 (forthcoming) and Harris & Anderson (2012), it is not only the course that matters, but also the levels at which students take a course.

[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.

[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. Because many students pursue the same course of study, only 3.7 percent of all students in this study are in singleton tracks; most students are in school tracks with more than 50 students, and the average student is in a school track with 117 other students.

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.

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*. To illustrate these sources of variation, consider the simple case illustrated in Table 4. There are two tracks, A and B, in a single school. There are two math teachers employed at the school at all times, but the identities of the teachers change from year to year due to staffing changes. The first source of variation is due to changes in staffing over time within schools. For example, between 2000 and 2005, Teacher 2 is replaced by Teacher 3. Because, teachers 2 and 3 both teach in track B (in different years), one can estimate the effect of Teacher 2 relative to Teacher 3 by comparing the outcomes of students with Teacher 2 in 2000 with those of students with Teacher 3 in 2005. To account for differences in outcomes between 2000 and 2005 that might confound comparisons within tracks over time (such as school-wide changes that may coincide with the hiring of new teachers), one can use the change in outcomes between 2000 and 2005 for Teacher 1 (who is in the school in both years) as a basis for comparison. In a regression setting this is accomplished with the inclusion of school-by-year fixed effects (Jackson & Bruegmann, 2009). This source of variation is valid as long as students do not select across cohorts (e.g., skip a grade) or schools in response to changes in Algebra I and English I teachers. Tests in section V provide little 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. In the example, because both teachers 1 and 2 taught students in track B in 2000, one can estimate the effect of Teacher 1 relative to Teacher 2 by comparing the outcomes of teachers 1 and 2 among students in track B in 2000. 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 V.2 show that the findings are not driven by student selection within school-track-years.[11]

[11] 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 and Algebra, 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.

IV.1 Estimating the Variance of Teacher Effects

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. As such, I follow Kane and Staiger (2008) and use the covariance between mean classroom-level residuals for the same teacher as my measure of the variance of teacher effects. This is done in two steps:

Step 1: Estimate equation [5] below.

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

There are no teacher (or classroom) indicator variables, 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 mean residuals from [5] for each classroom $\bar{e}_c^{*} \equiv \theta_j + \phi_c + \hat{\varepsilon}_c$ where $\hat{\varepsilon}_c$ is the classroom-level mean error term.

Step 2: 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. That is, compute $cov(\bar{e}_c^{*}, \bar{e}*_{c'} \mid J = j)$. If the classroom errors ϕ_{c} are uncorrelated with each other (recall that the model includes school-by-year fixed effects) 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 (Kane & Staiger, 2008). This is represented by $cov(\bar{e}_c^{*}, \bar{e}*_{c'} \mid J = j) = \mathrm{cov}(\theta_j, \theta_j) = \mathrm{var}(\theta_j) \longrightarrow \sigma^2_{\theta_j}$.[12] To ensure that the estimate is not driven by any particular random pairing of classrooms for the same teacher, I replicate this calculation 50 times and take the median of the estimated covariance as the parameter estimate. Following Jackson (forthcoming), I also compute bootstrap standard errors for the estimated covariance and use them for normal-distribution-based hypothesis testing and forming confidence intervals.[13]

V Main Results

V.1 True Variance of Teacher Effect on test score and non-test score outcomes

Table 5 presents the square root of the estimated covariance across classrooms for the same teachers. Because standard deviations are positive, when the sample covariance is negative (none

[12] Note that: $\mathrm{cov}(\theta_j,\phi_{c'}) = \mathrm{cov}(\theta_j,\bar{e}_{jgyc'}) = \mathrm{cov}(\phi_c,\theta_j) = \mathrm{cov}(\phi_c,\phi_{c'}) = \mathrm{cov}(\phi_c,\bar{e}_{jgyc'}) = \mathrm{cov}(\bar{e}_{jgyc},\theta_j) = \mathrm{cov}(\bar{e}_{jgyc},\phi_{c'}) = \mathrm{cov}(\bar{e}_{jgyc},\bar{e}_{jgyc'}) = 0$

[13] I use the standard deviation of 50 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.

of the negative covariance estimates is statistically significantly different from zero at the 5 percent level), I report the standard deviation to be zero. I present models that account for tracking with school-track fixed effects (top panel), and the preferred models that include both school-track fixed effects and school-by-year effects to account for both bias due to tracking and any school-wide shocks that might be confounded with teacher effects (lower panel). In models that include track-by-school fixed effects but not school-by-year effects (top panel), both algebra and English teachers had meaningful effects on test scores and non-test-score outcomes. Adding additional controls for school-by-year effects reduces the variability of the effects by about 30 percent.

In the preferred model (lower panel), the standard deviation of the algebra teacher effects on algebra test scores is 0.066σ (*p*-value of 0.000). The standard deviation of teacher effects is statistically significantly different from zero for some non-cognitive outcomes; the standard deviation of teacher effects on GPA is 0.045 grade points, and the effect on enrolling in 10th grade is 2.5 percentage points. The standard deviation of the effects on the non-cognitive factor is thus 0.083σ (*p*-value of 0.000). Looking to English I teachers, in the preferred model (right lower panel), the standard deviation of English teacher effects on English test scores is 0.034σ (*p*-value of 0.000). The estimated teacher effects are also statistically different from zero for all of the non-test score outcomes: the standard deviation of teacher effects on the likelihood of being suspended is 1.4 percentage points, the effect on absences is 3.7 percent, that on GPA is 0.027 grade points, and that on enrolling in 10th grade is 2.4 percentage points. Summarizing these effects, the standard deviation of English teacher effects on the non-cognitive factor is 0.071σ (*p*-value of 0.000).

To put the non-test score estimates into perspective, having an algebra or English teacher at the 85th percentile of effects on GPA versus the 15th percentile would be associated with 0.09 and 0.054 grade points higher GPA, respectively. For both subjects, having a teacher at the 85th percentile of effects on on-time grade progression (versus the 15th percentile) would be associated with being 5 percentage points (0.14σ) more likely to enroll in 10th grade on time. Students of an English teacher at the 85th percentile at reducing absences and suspensions versus the 15th percentile would be 2.8 percentage points (0.12σ) less likely to be suspended and have 7.4 percent fewer days absent. Teachers in both subjects had larger effects on the standardized non-cognitive factor than on standardized test scores. Overall, having an algebra teacher at the 85th percentile of improving non-cognitive ability versus the 15th percentile would be associated with 0.166σ higher non-cognitive ability, while the same calculation for English teachers is 0.142σ. Having established that teachers have real effects on both test scores and non-test score outcomes, Section V.3 investigates whether test score effects are correlated with effects on non-test-score outcomes. Section VI then investigates whether effects on non-test score outcomes predict effects on longer run outcomes above and beyond test score effects.

V.2 Tests for Bias due to Selection

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 school-tracks.[14] I show that this condition is likely satisfied for algebra teachers at all schools, and for English teachers at most schools. Accordingly, in the following section I focus on algebra teachers at all schools and English teachers at this subsample of schools. Using tests for student selection to teachers on observable dimensions and also unobservable dimensions within school-tracks, I show that there is little evidence of selection bias.

To test for selection on observables within school tracks, I follow Chetty, Friedman, and Rockoff (2011). I predict each outcome (based on 7th and 8th grade math and reading scores, parental education, gender, and ethnicity) and regress predicted outcomes on school-track indicators, year indicators, and the estimated effect of the student's teacher (estimated out of sample).[15] If students with characteristics associated with better outcomes selected to classrooms based on teacher effectiveness, then there would be a systematic relationship between estimated teacher quality and predicted outcomes. The results (lower panel of Table 6) indicate that where there are multiple algebra teachers in the same school track, algebra teachers with higher estimated effects do not receive students with better or worse predicted outcomes on average—suggesting no selection of algebra teachers within tracks.

The same test for English teachers suggests no selection to teachers based on the effects on the non-cognitive factor (column 4), but possibly some positive selection based on test score value-added (column 3). If this positive selection for English teachers exists to some degree at all schools,

[14] 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).

[15] To remove any endogeneity, for each observation year I estimate teacher effects using all *other* years of data. For example, for observations in 2005, the estimated teacher effects are based on teacher performance in 2006, 2007, 2008, 2009, and 2010. For estimates in 2008, estimates are based on 2005, 2006, 2007, 2009, and 2010. I follow Kane and Staiger (2008) and compute a teacher's fixed effects using an efficiently weighted average of all the teacher's mean classroom level residuals (See Appendix Note 2 for details).

then it would invalidate inferences regarding English teachers. However, if these average effects are driven by a few outlier schools (i.e., there is positive selection within tracks only at a minority of schools), because all inferences are based on within-school comparisons one can obtain valid inference by removing those schools that exhibit positive selection. To test for this possibility, I conducted the selection test for each school and found that while most schools do not exhibit significant selection to English teachers within tracks, a small number of schools do. Accordingly, I removed those schools that exhibit evidence of positive selection to teachers within tracks to create a "clean" English teacher sample.[16] *Note that because all estimates are obtained within schools, removing entire schools that may exhibit selection within tracks does not introduce sample selection bias or endogeneity.* While this reduces the sample size for English teachers by roughly 9 percent, within this clean sample, there is no evidence of student selection within tracks (columns 5 and 6). I use this clean sample for analysis of English teachers from this point forward.

To test for selection on unobservables within school-track-years, 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, then the arrival of a teacher with a high positive estimated effect (who increases the average estimated teacher effect for a cohort) should have no effect on average student outcomes for that cohort. Conversely, if the estimated effects are real, differences in average estimated 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 [6] and [7], where $\hat{\theta}_j$ is the estimated (out of sample) effect of teacher *j*, $\bar{\hat{\theta}}_{j\in sgy}$ is the mean estimated 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.

$$Y_{isgjcy} = A_{iy-1}\delta + \psi_1\hat{\theta}_j + X_{iy}\beta + \theta_{\mathrm{sgy}} + \varepsilon_{isgjcy} \qquad [6]$$

[16] I regress predicted English scores on the teacher's estimated value-added, school-track fixed effects, and year effects for each school. Any school that yields a t-statistic greater than 2.5 on teacher value-added is dropped from the analytic sample.

$$Y_{iisgjcy} = A_{iy-1}\delta + \psi_2 \bar{\hat{\theta}}_{j \in sgy} + X_{iy}\beta + I_{sgi}\,\theta_{\mathrm{sg}} + \theta_{\mathrm{sy}} + \varepsilon_{iisgjcy} \qquad [7]$$

Equations [6] and [7] both calculate estimated teacher effects on student outcomes, but each use a distinct source of variation. In [6], 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 [7], 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.

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 teacher effects are estimated with noise, the coefficients ψ_1 and ψ_2 will be less than 1. The results are presented in the top panel of Table 6. Despite there being no relationship between estimated teacher quality and *predicted* outcomes, there are economically and statistically significant effects of estimated teacher quality on *actual* outcomes for both subjects. Marginal effects obtained using variation within school-track-cohorts are similar to those obtained using variation across cohorts within school-tracks. Also, mean school-track cohort-level teacher quality has a statistically significant effect on all outcomes, so the null hypothesis that estimated teacher effects are driven by selection within school-track-cohorts can be rejected at the 5 percent level.

The discussion thus far has focused on selection within school-tracks. However, readers might wonder if the results are biased due to student selection across tracks. To test for this, I regress student outcomes on the school-year level mean estimated 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 6 show that mean school cohort level teacher quality has a statistically significant effect on all outcomes and is similar to those from teacher level variation – indicating that selection across tracks does not bias the results.

V.3 *Relationship between Teacher Effects across Outcomes*

Having established that teachers have real causal effects on test scores, non-test-score outcomes, and a proxy for non-cognitive ability, 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 estimated 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 7 measures the fraction of the 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 value-added are associated with better non-test-score outcomes, but the relationships are weak. Effects on algebra test scores explain 1.15 percent of the variance in estimated teacher effects on suspensions, 2.09 percent for absences, 9.84 percent for GPA, and 4.97 percent of the effect 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. In contrast, effects on the non-cognitive factor explain much of the estimated effects on the non-test score outcomes. Specifically, algebra teacher effects on the non-cognitive factor explain 31.6 percent of the estimated teacher effect on suspensions, 37.8 percent for absences, 62.6 percent for GPA, and 62.07 percent of the effect on on-time 10th grade enrollment (top panel second row). However, teacher effects on the non-cognitive factor explain less than 10 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 10 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). However, English teacher effects on the non-cognitive factor explain 30.18 percent of the variance in teacher effect on suspensions, 39.8 percent for absences, 61.86 percent for GPA, and 68.09 percent of the effect on on-time 10th grade enrollment.

In sum, for both subjects, teacher test-score effects measure certain skills, and teacher effects on the non-cognitive factor measure a largely *different* but also important set of skills. For both subjects, a teacher's effect on test scores is a weak predictor of her effect on the non-cognitive factor. To show this visually, Figure 2 presents a scatterplot of teachers' estimated effects on the non-cognitive factor against their effect on test scores. It is clear that teacher effects on test scores in both subjects may leave much variability in effects on non-cognitive skills unexplained. This indicates that many teachers who improve test scores may have average effects on non-test-score outcomes. Similarly, 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 likely 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. How this affects our ability to identify excellent teachers depends on whether teacher effects on the non-cognitive factor provide more information on their effectiveness at improving longer-run outcomes than that conveyed by their effects on test scores. Section VI investigates this.

VI Predicting Long 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, it is important to show that teachers who increase the non-cognitive factor actually *cause* students to have improved long-run outcomes (conditional on their test score effects). To test this, I link estimated teacher effects (estimated out of sample) to variables denoting whether the student subsequently dropped out of secondary school, graduated from high school, took the SAT, or expressed plans to attend college. 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 estimated (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.

$$Y_{ijcy} = A_{iy-1}\delta + \psi_{1,test}\hat{\theta}_{j,test} + X_{iy}\beta + I_{sgi}\,\theta_{\text{sg}} + \theta_{\text{sy}} + \varepsilon_{ijcy} \qquad [8]$$

$$Y_{ijcy} = A_{iy-1}\delta + \psi_{2,test}\hat{\theta}_{j,test} + \psi_{2,noncog}\hat{\theta}_{j,noncog} + X_{iy}\beta + I_{sgi}\,\theta_{\text{sg}} + \theta_{\text{sy}} + \varepsilon_{ijcy} \qquad [9]$$

To quantify the extent to which including both $\hat{\theta}_{j,noncog}$ and $\hat{\theta}_{j,test}$ in [9] increases our ability to predict variability in teacher effects over only including $\hat{\theta}_{j,test}$ in [8], I computed the percentage increase in the predicted variability of the teacher effects on the long-run outcome from [8] to [9]. I computed $100\times\left(sd(\hat{\psi}_{2,test}\hat{\theta}_{j,test} + \hat{\psi}_{2,noncog}\hat{\theta}_{j,noncog}) / sd(\hat{\psi}_{1,test}\hat{\theta}_{j,test}) - 1\right)$.

Column 1 of Table 8 shows that students with algebra teachers who raise test scores by 1σ are 0.327 percentage points less likely to drop out of high school. While this effect has the expected positive sign, the magnitude is small and the point estimate is not statistically significantly different from zero. Column 2 shows that while a teacher's effect on test scores has little association with dropout, teacher effects on the non-cognitive factor have a statistically significant negative relationship with dropout. Students with an algebra teacher who raises the non-cognitive factor by 1σ are 1.05 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 430 percent. Because the standard deviation of teacher effects on the non-cognitive factor is roughly 0.08, going from a teacher at the 15th to one at the 85th percentile of the non-cognitive effect distribution is associated with 0.168 percentage points less chance of dropping out. While this effect may seem small, small 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). Though they do not affect dropout, teacher effects on algebra scores do predict greater high school graduation. Students with teachers that raise test scores by 1σ are 2.16 percentage points more likely to graduate high school. Adding teacher effects on the non-cognitive factor reduces the coefficient on teacher test score effects (reflecting that they are positively correlated) and increases the standard deviation of the predicted effect by about one-third. Having a teacher at the 85th as opposed to the 15th percentile of both the non-cognitive effect distribution and the test score effect distribution is associated with being 0.36 percentage points more likely to graduate from high-school. Results in columns 5 through 8 suggest that algebra teachers have little effect on whether students take the SAT, but do effect whether students plan to attend college (reported at high school graduation). Students with teachers that raise test scores by 1σ are 3.5 percentage points more likely to have plans to attend college. This is statistically indistinguishable from the Chetty et al. (2011) estimate of teacher test score effects on the likelihood of college going of 4.9 (se=0.65). Adding teacher effects on the non-cognitive factor increases the predicted standard deviation by 22.5 percent. In sum, the results suggest that algebra teacher effects on test scores do predict their effects on longer run outcomes, but that adding teacher effects on the non-cognitive factor increases the predicted variability by about one-third for high school graduation and indicators of college going, and increases the explained variability of effects on dropout by about 430 percent.

The results for English teachers (Table 9) suggest even greater impact of teacher effects on non-cognitive skills. Column 1 shows that students who have teachers that raise English test scores by 1σ are 1.14 percentage points less likely to drop out of high school. Similar to algebra, Column 2 shows that a teacher's effect on the non-cognitive factor has a statistically significant negative relationship with dropout (conditional on test score effects), and increases the variability in the explained teacher effects on dropout by 115 percent. Furthermore, students with English teachers that raise test scores by 1σ are just 1.9 percentage points less likely to drop out of high school. However, teacher effects on the non-cognitive factor have a positive and statistically significant relationship with graduating high school (conditional on test score effects), and increase the predicted variability by 204 percent. Overall, going from a teacher at the 15th percentile to one at the 85th of both the non-cognitive factor and the test score distribution is associated with being 0.35 percentage points more likely to graduate from high-school. Results in columns 5 and 6 suggest that English teachers have important effects on whether a student takes the SAT and plans to attend college. Students with teachers that raise English scores by 1σ are 6.86 percentage points more likely to take the SAT. The non-cognitive factor also has a positive and statistically significant relationship with SAT taking (conditional on test score effects), and increases the variability in the explained teacher effect by 51 percent. Going from a teacher at the 15th to one at the 85th percentile of both the non-cognitive effect distributions is associated with being 0.77 percentage points more likely to take the SAT. Finally, column 7 shows that students with teachers that raise English scores by 1σ are 0.7 percentage points more likely to plan to attend college. Adding effects on the non-cognitive factor increases the predicted standard deviation by 697 percent, so that going from a teacher at the 15th to one at the 85th percentile of both the non-cognitive and the test-score effect distributions is associated with being 0.4 percentage points more likely to plan to attend college.[17]

Overall, the results suggest that both algebra and English teachers' effects on test scores predict their effects on students' long-run outcomes. However, the results also show that teacher effects on the non-cognitive factor affect long-run outcomes in a statistically significant way, conditional on teacher effects on test scores. As teacher effects on the non-cognitive factor may capture their effects on important skills unmeasured by test scores, it follows that adding teacher

[17] The fact that teacher effects on the non-cognitive factor are greater in English than Algebra is consistent with Chetty et al.'s (2011) finding that marginal increases in English teacher quality have larger effects on longer-run outcomes, even though effects on English tests are smaller than those for math.

effects on the non-cognitive factor increases the predicted variability on longer-run outcomes by between 30 and 700 percent (depending on the outcome and the subject teacher). To ensure that these estimated effects are not due to student selection, the lower panels of Table 8 and 9 show the estimated effect on *predicted* longer run outcomes (based on all observable covariates) with only school-track fixed effects and school-year fixed effects. There is no relationship between teacher effects and students' *predicted* outcomes—indicating that these results can reasonably be interpreted causally. Note 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 *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 correlated with the test score effect and may explain variability in the long run effect unexplained by value-added.[18] Accordingly, it is important to know if the non-cognitive factor truly measures a different set of skills than test scores do, 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

[18] 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.

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 estimated effects on the non-cognitive factor and out of sample estimated effects on test scores (with school-track fixed effects and school-by-year fixed effects). Conditional on test score effects, teacher effects on the non-cognitive factor yield a *p*-values of 0.291 for algebra test scores and a *p*-value 0.73 for English test scores. That is, teacher effects on the non-cognitive factor provide no additional predictive power for test scores. This is inconsistent with measurement error in value-added causing effects on the non-cognitive factor to explain effects on long run outcomes. Accordingly, the results suggest that long-run effects reflect multiple dimensions of skills and that the non-cognitive factor captures dimensions of ability not 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 3. 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 3 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.[19]

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 and 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.

[19] 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 and Lockwood (2013).

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 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. The results indicate that adding teacher effects on the non-cognitive factor increases the predicted variability on longer-run outcomes by between 30 and 700 percent.

The findings suggest that test-score measures understate the effect of teachers on adult outcome in general, and may greatly understate their importance in affecting outcomes that are heavily determined by non-cognitive skills (such as dropping out and criminality). While the results are not entirely surprising, they do provide the first evidence that measuring teacher effects on test scores captures only a fraction of their effect on longer-run outcomes. They also suggest that evaluating teacher effects on non-test-score outcomes may greatly improve our ability to predict teachers' overall effects on longer-run outcomes. 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. 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. 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.
5. 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.
6. Borghans, L., Weel, B. T., & Weinberg, B. A. (2008). Interpersonal Styles and Labor Market Outcomes. *Journal of Human Resources, 43*(4), 815-58.
7. Bowles, S., Gintis, H., & Osborne, M. (2001). The Determinants of Earnings: A Behavioral Approach. *Behavioral Approach, 39*(4), 1137-76.
8. Brookhart, S. M. (1993). Teachers' Grading Practices: Meaning and Values. *Journal of Educational Measurement, 30*(2), 123-142.
9. Carneiro, P., Crawford, C., & Goodman, A. (2007). The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes. *CEE Discussion Papers 0092*.
10. Cascio, E., & Staiger, D. (2012). Knowledge, Tests, and Fadeout in Educational Interventions. *NBER working Paper Number 18038*.
11. 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.
12. Chetty, R., Friedman, J., & Rockoff, J. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. (Unpublished manuscript).
13. Deming. (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111-134.
14. Deming, D. (2011). Better Schools, Less Crime? *The Quarterly Journal of Economics, 126*(4), 2063-2115.
15. Downey, D., & Shana., P. (2004). When Race Matters: Teachers' Evaluations of Students' Classroom Behavior. *Sociology of Education, 77*, 267-82.
16. 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.
17. 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.
18. 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.
19. Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica, 77*(3), 953-973.
20. Fredriksson, P., Ockert, B., & Oosterbeek, H. (forthcoming). Long-Term Effects of Class Size. *Quartlerly Journal of Economics*.
21. 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.
22. 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.
23. Heckman, J. (1999). Policies to Foster Human Capital. *NBER Working Paper 7288*.
24. Heckman, J. J., & Rubinstein, Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review, 91*(2), 145-49.
25. 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.
26. Heckman, J., Pinto, R., & Savelyev, P. (forthcoming). Understanding the Mechanisms Through Which an

Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*.

27. 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.
28. Howley, A., Kusimo, P. S., & Parrott, L. (2000). Grading and the ethos of effort. *Learning Environments Research, 3*, 229-246.
29. Jackson, C. K. (forthcoming). Teacher Quality at the High-School Level: The Importance of Accounting for Tracks. *Journal of Labor Economics*.
30. Jackson, C. K. (forthcoming). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics.*
31. 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.
32. Jencks, C. (1979). *Who Gets Ahead? The Determinants of Economic Success in America.* New York: Basic Books.
33. 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.
34. 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.
35. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Pearson.
36. Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER working paper 14607*.
37. Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER Working Paper # 14607.*
38. Kinsler, J. (2012). Assessing Rothstein’s critique of teacher value-addedmodels. *Quantitative Economics, 3*, 333-362.
39. Koedel, C. (2008). An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Department of Economics, University of Missouri Working Paper 0808*.
40. Koedel, C. (2008). Teacher Quality and Dropout Outcomes in a Large, Urban School District. *Journal of Urban Economics, 64*(3), 560-572.
41. Koedel, C., & Betts, J. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? *Education Finance and Policy, 6*(1), 18-42.
42. 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.
43. 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.
44. Lucas, S. R., & Berends, M. (2002). Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking. *Sociology of Education, 75*(4), 328-348.
45. Mihaly, Kata., Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood (2013). "A Composite Estimator of Effective Teaching" Gates foundation Research Paper.
46. Mansfield, R. (2012). Teacher Quality and Student Inequality. (Working Paper) *Cornell University*.
47. Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73*(2), 417-458.
48. Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*.
49. Sadker, D. M., & Zittleman, K. (2006). *Teachers, Schools and Society: A Brief Introduction to Education.* McGraw-Hill.
50. Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance."*American Economic Review*, 102(7): 3628-51.
51. 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	0.091	(0.944)	(0.600)	(0.878)
Reading z-score 8th grade	0.073	(0.941)	(0.678)	(0.891)
Male	0.510	(0.50)	(0.482)	(0.498)
Black	0.288	(0.453)	(0.375)	(0.399)
Hispanic	0.075	(0.263)	(0.245)	(0.256)
White	0.579	(0.494)	(0.404)	(0.432)
Asian	0.020	(0.141)	(0.133)	(0.138)
Parental education: Some High-school	0.075	(0.263)	(0.25)	(0.259)
Parental education: High-school Grad	0.400	(0.49)	(0.454)	(0.474)
Parental education: Trade School Grad	0.018	(0.132)	(0.129)	(0.132)
Parental education: Community College Grad	0.133	(0.339)	(0.327)	(0.335)
Parental education: Four-year College Grad	0.205	(0.404)	(0.376)	(0.394)
Parental education: Graduate School Grad	0.064	(0.245)	(0.225)	(0.237)
Number of Honors classes	0.880	(1.323)	(0.575)	(1.163)
Algebra I z-Score (9th grade)	0.063	(0.976)	(0.775)	(0.889)
English I z-Score (9th grade)	0.033	(0.957)	(0.670)	(0.906)
Ln Absences	0.586	(1.149)	(0.927)	(0.984)
Suspended	0.056	(0.23)	(0.214)	(0.225)
GPA	2.763	(0.87)	(0.604)	(0.801)
In 10th grade	0.856	(0.351)	(0.305)	(0.339)
Dropout (2005-2010 cohorts)	0.083	(0.276)	(0.205)	(0.213)
Graduate (2005-2009 cohorts)	0.793	(0.405)	(0.380)	(0.405)
Take SAT (2005-2009 cohorts)	0.393	(0.489)	(0.386)	(0.439)
Intend to attend college (2005-2009 cohorts)	0.387	(0.487)	(0.432)	(0.463)
Observations			348547	

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. About 10 percent of students do not have parental education data—the missing category is “missing parental education”.

Table 2: *Correlations between the short run outcomes*

	Raw correlations between outcomes						Percentage of Variance Explained by Factors		
	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
Ln of # Days Absent	1						0.010	0.007	0.281
Suspended	0.252	1					0.017	0.017	0.365
Grade Point Average	-0.232	-0.192	1				0.350	0.291	0.677
In 10th grade on time	-0.167	-0.16	0.482	1			0.096	0.095	0.563
Algebra Score in 9th Grade	-0.098	-0.13	0.592	0.310	1		1.000	0.379	0.234
English Score in 9th Grade	-0.082	-0.13	0.539	0.308	0.616	1	0.379	1.000	0.271

Note: The cognitive and non-cognitive factors were uncovered using factor analysis and are linear combinations of all the short-run outcomes. The results were then standardized. Note that the factors in the NCERDC use the weights derived from the NELS-88 data. However, the factors using weights derived from the NCERDC have correlations greater than 0.95 with those derived using weights from the NELS-88.

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

	1	2	3	4	5	6
	Dataset: NCERDC Micro Data					
	Drop out	Graduate	Take SAT	Drop out	Graduate	Take SAT
Grade Point Average	-0.0360**	0.116**	0.187**			
	[0.00118]	[0.00217]	[0.00214]			
Log of # Absences	0.00814**	-0.0284**	-0.0449**			
	[0.000522]	[0.00104]	[0.00122]			
Suspended	0.0128**	-0.0505**	-0.0160**			
	[0.00311]	[0.00572]	[0.00474]			
On time in 10th grade	-0.0755**	0.290**	0.0374**			
	[0.00260]	[0.00454]	[0.00329]			
English z-score	-0.00501**	0.00461*	0.0185**			
	[0.00111]	[0.00203]	[0.00215]			
Math z-score	-0.00880**	0.0192**	0.0256**	-0.00710**	0.0183**	0.0889**
	[0.000904]	[0.00170]	[0.00180]	[0.000572]	[0.00117]	[0.00120]
Non-cog factor z-score				-0.0485**	0.182**	0.152**
				[0.000846]	[0.00137]	[0.00115]
School Fixed Effects	Y	Y	Y	Y	Y	Y
Covariates	Y	Y	Y	Y	Y	Y
Observations	208,330	171,226	171,226	208,330	171,226	171,226

Robust standard errors in brackets. ** p<0.01, * p<0.05, + p<0.1
All models include controls for student gender, ethnicity, are parental education.

Table 4: *Illustration of the Variation at a Hypothetical School*

		Track A	Track B
		Alg I (regular) Eng I (regular) Natural Sciences US History	Alg I (regular) Eng I (regular) Biology World History Geometry
	Year		
Math Teacher 1	2000	X	X
Math Teacher 2	2000		X
Math Teacher 1	2005	X	X
Math Teacher 2*	2005	-	-
Math Teacher 3	2005		X

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

		Algebra Teachers				English Teachers			
		SD	Prob Cov≤0	95% CI Upper bound	95% CI Lower bound	SD	Prob Cov≤0	95% CI Upper bound	95% CI Lower bound
Track-by-School and Year Effects	Algebra Score 9th	0.096	0.000	0.107	0.093	0.042	0.005	0.051	0.019
	English Score 9th	0.013	0.842	0.034	0.000	0.043	0.000	0.050	0.035
	Suspended	0.009	0.645	0.022	0.000	0.021	0.000	0.025	0.017
	Log of # Absences	0.093	0.000	0.109	0.073	0.095	0.000	0.108	0.08
	GPA	0.065	0.000	0.075	0.053	0.049	0.000	0.059	0.037
	On time enrollment	0.035	0.000	0.042	0.027	0.031	0.000	0.037	0.025
	Non-cognitive factor	0.113	0.000	0.127	0.097	0.092	0.000	0.103	0.08
Track-by-School and School-by-Year Effects	Algebra Score 9th	0.066	0.000	0.074	0.056	0.000	0.968	0.008	0.000
	English Score 9th	0.000	0.917	0.009	0.000	0.034	0.000	0.041	0.025
	Suspended	0.000	0.887	0.008	0.000	0.014	0.003	0.019	0.007
	Log of # Absences	0.000	0.798	0.03	0.000	0.037	0.024	0.054	0.009
	GPA	0.045	0.000	0.057	0.028	0.027	0.007	0.039	0.008
	On time enrollment	0.025	0.002	0.033	0.012	0.024	0.000	0.031	0.015
	Non-cognitive factor	0.083	0.000	0.06	0.100	0.071	0.000	0.082	0.056

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 50 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 50 replications of these "placebo" covariances is my bootstrap estimate of the standard deviation of the estimated covariance. These two estimates can then be 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). When the estimated covariance is negative, I report a value of zero for the standard deviation. Note that none of the negative covariances are statistically significant at the five percent level.

Table 6: *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	5	6
	Algebra Teachers		English Teachers		English Teachers (clean sample) b	
	Algebra	Non-cognitive	English	Non-cognitive	English	Non-cognitive
Estimated Effect (*within cohorts*)	0.275**	0.0599**	0.197**	0.0732**	0.196**	0.0733**
	[0.0350]	[0.0198]	[0.0235]	[0.0191]	[0.0255]	[0.0203]
Mean Estimated Effect (*across cohorts*)	0.263**	0.103*	0.272**	0.168**	0.267**	0.178**
	[0.0612]	[0.0495]	[0.0444]	[0.0341]	[0.0460]	[0.0357]
School-year Mean Effect (*across tracks*)	0.405**	0.0641*	-	-	0.130**	0.112**
	[0.0754]	[0.0317]	-	-	[0.0453]	[0.0365]
	Predicted Algebra	Predicted Non-cognitive	Predicted English	Predicted Non-cognitive	Predicted English	Predicted Non-cognitive
Estimated Effect (*all variation*) [a]	0.019	-0.00376	0.0842**	-0.00718	0.0224	-0.0114
	[0.0223]	[0.00355]	[0.0247]	[0.00622]	[0.0262]	[0.00695]
Observations	137,600	139,173	284,363	284,363	256,308	256,308

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

All models include school-year effects and school-track fixed effects. The independent variable in-within cohort models is the estimated 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. Note: the predicted outcome reflects the effects of 7th and 8th grade test scores, parental education, gender, and ethnicity.

b. The clean English sample removes those schools that presented strong statistical evidence of non-random sorting of students to teachers within tracks. To do this, I regressed predicted English scores on out-of-sample teacher value-added for each school. Any school for which the t-statistic on teacher value added was larger than 2.5 were removed from the sample.

Table 7: *Proportion of the Variability in Estimated Effects Explained by Estimated Effects on Test Scores and Effects on the Non-cognitive Factor**

	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 effect	**1**	-	0.0115	0.0209	0.0984	0.0497	0.091
Non-cognitive factor effect	0.1	-	**0.3165**	**0.3781**	**0.6267**	**0.6207**	**1**
English Test score effect	-	1	0.0116	0.0247	0.0534	0.0585	0.0483
Non-cognitive factor effect	-	0.0730	**0.3018**	**0.3980**	**0.6186**	**0.6809**	1

*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 8: *Effect of Short-run Algebra Teacher effects on Long-run Outcomes*

	1	2	3	4	5	6	7	8
	Dropout	Dropout	Graduate	Graduate	Take SAT	Take SAT	Intend college	intend college
Effect on test score	-0.00327	0.0062	0.0216*	0.0127	0.0117	0.0117	0.0350*	0.0266+
	[0.00633]	[0.0058]	[0.0104]	[0.0112]	[0.0130]	[0.0141]	[0.0145]	[0.0138]
Effect on non-cog factor		-0.0105*		0.0106		4.96E-05		0.00976
		[0.0042]		[0.0083]		[0.0082]		[0.0087]
% increase sd(θ*)	430.86		32.93		6.4		22.5	
Observations	139203		113939		113939		99640	

	Predicted Dropout	Predicted Graduate	Predicted SAT	Predicted Intend
Effect on test score	-0.00099	0.0034	0.00847+	0.00431
	[0.000768]	[0.00249]	[0.00450]	[0.00264]
Effect on noncog factor	0.000261	-0.00076	-0.00182	-0.00116
	[0.000425]	[0.00141]	[0.00272]	[0.00206]

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

Note: % increase sd(θ*) is the percentage increase in the standard deviation of the fitted values associated with estimated teacher-added.

Table 9: *Effect of Short-run English Teacher effects on Long-run Outcomes (clean sample)*

	1	2	3	4	5	6	7	8
	Dropout	Dropout	Graduate	Graduate	Take SAT	Take SAT	Intend college	Intend college
Effect on test score	-0.0114	-0.0018	0.019	-0.00235	0.0686**	0.0394*	0.00738	-0.0163
	[0.0091]	[0.0102]	[0.0164]	[0.0169]	[0.0147]	[0.0162]	[0.0179]	[0.0188]
Effect on non-cog factor		-0.0110*		0.0261**		0.0358**		0.0286**
		[0.0045]		[0.0073]		[0.0078]		[0.0091]
% increase sd(θ*)	115.25		204.24		51.29		697.86	
Observations	258,706		220,706		220,706		181,762	

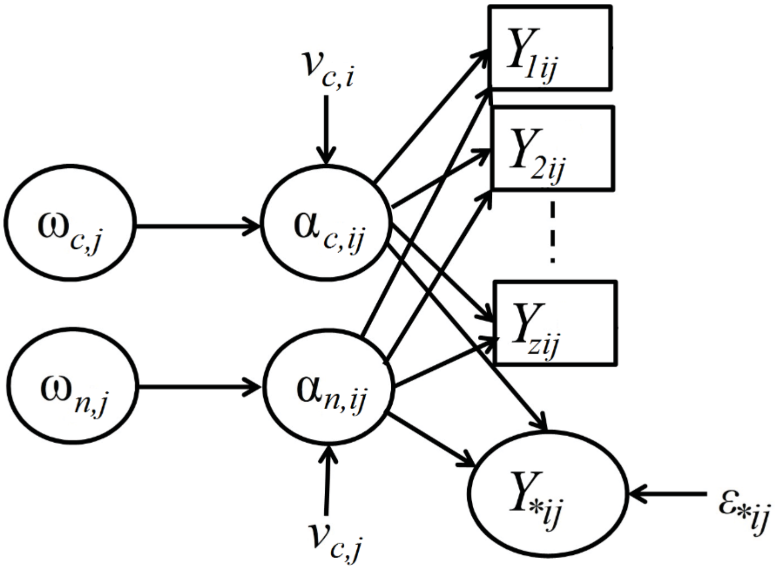
	Predicted Dropout	Predicted Graduate	Predicted SAT	Predicted Intend
Effect on test score	1.95E-05	-2.10E-04	8.60E-03	2.05E-03
	[0.00182]	[0.00504]	[0.00656]	[0.00751]
Effect on noncog factor	0.00183	-0.00376*	-0.00183	-0.00324
	[0.001201]	[0.00178]	[0.00257]	[0.00243]

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

Note: % increase sd(θ*) is the percentage increase in the standard deviation of the fitted values associated with estimated teacher-added.

Figures

Figure 1: *Path Diagram of the Two-Factor Model*



Note: An arrow from *a* to *b* indicates that variable *b* is a linear function of variable *a*. Square boxes denote observed variables; while ovals denote unobserved or latent variables.

Figure 2: *Relationship between Teacher Effects on Test Scores and Non-cognitive Factor*

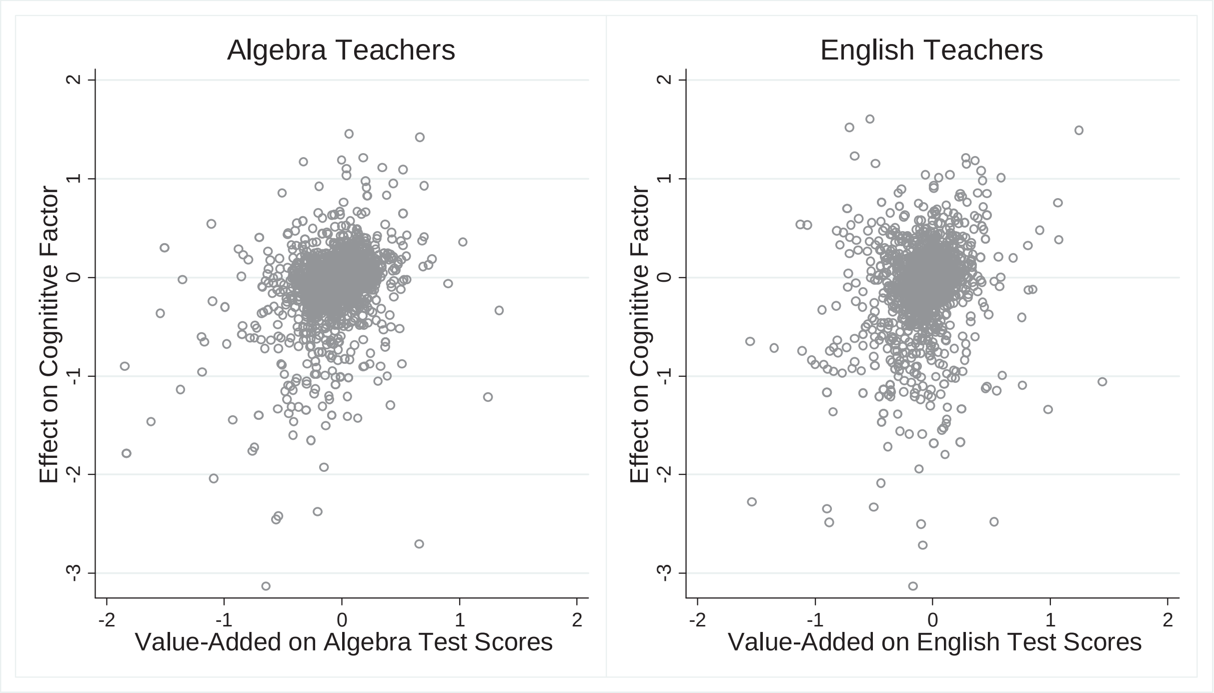
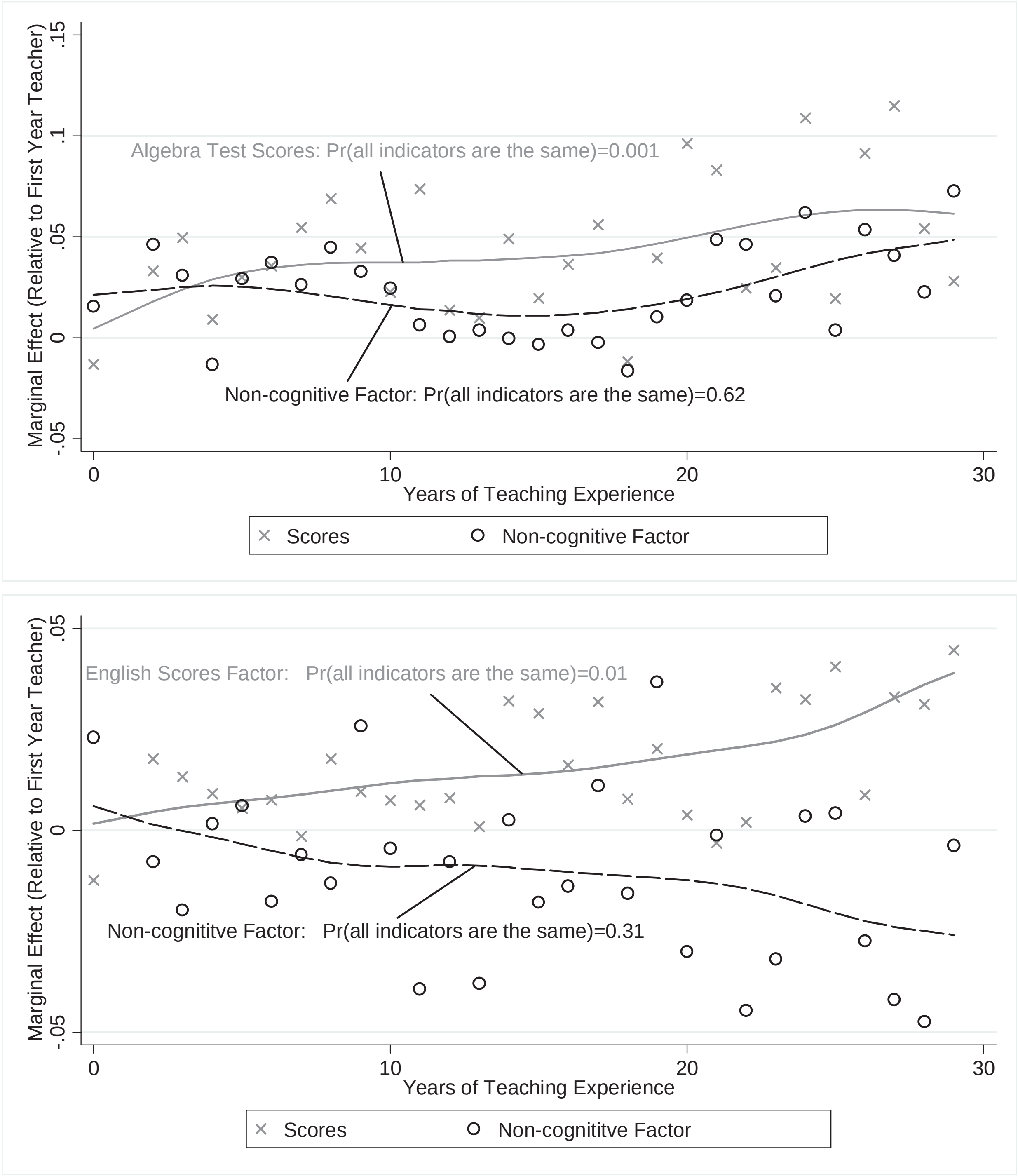


Figure 3: *Effect of Experience on Test Scores and Non-cognitive Factor*



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 60 percent of classrooms. All 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 will tend to have more precise estimates and (2) there are classroom level disturbances so that teachers with multiple classrooms will have more precise estimates. As before, I compute mean residuals from [7] for each classroom $\overline{e}^{*}_{\mathrm{cj}} \equiv \theta_j + \phi_c + \hat{\varepsilon}_c$. Since the classroom error is randomly distributed, I use the covariance between the mean residuals of classrooms for the same teacher $\operatorname{cov}(\overline{e}^{*}_{cj}, \overline{e}^{*}_{c'j}) = \hat{\sigma}^2_{\theta_j}$ 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 high 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), thereby 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