Institute for Policy Research
Northwestern University
Working Paper Series

WP-12-18

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

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

Version: December 2012

DRAFT
Please do not quote or distribute without permission.

2040 Sheridan Rd. ◆ Evanston, IL 60208-4100 ◆ Tel: 847-491-3395 Fax: 847-491-9916
www.ipr.northwestern.edu ◆ ipr@northwestern.edu

Abstract

In this working paper, Jackson presents a model where students have cognitive and non-cognitive ability and a teacher's effect on long-run outcomes is a combination of his or her effect on both ability types. Conditional on cognitive scores, an underlying non-cognitive factor associated with student absences, suspensions, grades, and grade progression is strongly correlated with long-run educational attainment, arrests, and earnings in survey data. In administrative data, teachers have meaningful causal effects on both test scores and the non-cognitive factor. Calculations indicate that teacher effects based on test scores alone fail to identify many excellent teachers, and may greatly understate the importance of teachers on adult outcomes.

.

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

C. Kirabo Jackson
Northwestern University, December 4, 2012

I present a model where students have cognitive and non-cognitive ability and a teacher's effect on long-run outcomes is a combination of her effect on both ability types. Conditional on cognitive scores, an underlying non-cognitive factor associated with student absences, suspensions, grades, and grade progression, is strongly correlated with long-run educational attainment, arrests, and earnings in survey data. In administrative data teachers have meaningful causal effects on both test-scores and this non-cognitive factor. Calculations indicate that teacher effects based on test scores alone fail to identify many excellent teachers, and may greatly understate the importance of teachers on adult outcomes. (JEL I21, J00)

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

There is a growing consensus that non-cognitive skills not captured by standardized tests, such as adaptability, self-restraint, self-efficacy, and motivation are important determinants of adult outcomes (Lindqvist & Vestman, 2011; Heckman & Rubinstein, 2001; Borghans, Weel, & Weinberg, 2008; Waddell, 2006). At the same time, many interventions that have no effect on test-scores have meaningful effects on long-run outcomes such as education, earnings, and crime (Booker, Sass, Gill, & Zimmer, 2011; Deming, 2009; Deming, 2011). Also Heckman, Pinto, & Savelyev (forthcoming) find that changes in personality traits explain the positive effect of the Perry Preschool Program on adult outcomes, and Fredriksson, Ockert, & Oosterbeek, (forthcoming) find that class size effects on earnings are predicted by effects on *both* test scores and non-cognitive ability. This suggests that schooling inputs produce both cognitive skills (measured by standardized tests) and non-cognitive skills (reflected in socio-behavioral outcomes), both of which determine adult outcomes. Accordingly, evaluating interventions based on test scores captures only one dimension of the skills required for adult success, and "*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).

[1] I thank David Figlio, Jon Guryan, Simone Ispa-Landa, Clement Jackson, Mike Lovenheim, and James Pustejovsky for insightful comments. I also thank Kara Bonneau from the NCERDC and Shayna Silverstein.

There is broad consensus among policy makers, educators, parents, and researchers that teachers are one of the most important components of the schooling environment. Studies show that having a teacher at the 85th percentile of the quality distribution (as measured by effects on student test scores) versus one at the 15th percentile is associated with between 8 and 20 percentile points higher test-scores in math and reading (Aaronson, Barrow, & Sander, 2007; Kane & Staiger, 2008; Chetty, Friedman, & Rockoff, 2011; Rivkin, Hanushek, & Kain, 2005). Several districts, including New York, D.C., and Los Angeles, 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 a teacher's effect on long-run outcomes and not *only* her effect on cognitive ability. To shed light on this issue, I test whether teachers have causal effects on *both* cognitive ability (measured by test scores) and non-cognitive ability (measured by absences, suspensions, grades, and grade progression). I also investigate whether teachers who improve test scores also improve non-test score outcomes, and estimate the extent to which test score measures understate the overall importance of teachers. While there is a growing literature on the importance of non-cognitive skills, and a burgeoning literature on the effect of teachers on test scores, this paper presents the first comprehensive analysis of teacher effects on *both* cognitive and objective non-cognitive outcomes.

Opponents of using test-scores to infer teacher quality raise two concerns. The first concern is that improving test scores does not necessarily imply better long-run outcomes because teachers might teach to the test, test score improvements might reflect transitory student effort, and those skills measured by test-scores may not be those that are associated with improved long-term outcomes. However, Chetty, Friedman, & Rockoff (2011) assuage this concern by demonstrating that teachers who improve test scores also improve 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 captured by test scores to increase test score outcomes – potentially decreasing teacher quality overall (Holmstrom & Milgrom, 1991). I speak to this second critique by assessing whether teachers affect non-cognitive skills not captured by test scores.

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 use teacher evaluations of students and may reflect differences in teacher perception rather than actual student behavior.[2] Koedel (2008) uses a design that assumes that students' "math-course choices are exogenous to their dropout decisions" to estimate high school teacher effects on graduation. However, both Harris & Anderson (2012) and Jackson (2012) document severe biases associated with tracking in middle- and high school so any estimated teacher effects may be spurious. Accordingly, the extent to which teachers affect non-cognitive outcomes remains unclear.

This paper is organized into three sections. In the first section, I present a latent factor model following Heckman, Stixrud, & Urzua (2006) in which both student ability and teacher ability have a cognitive and a non-cognitive dimension. I show that a teacher's effect on long-run outcomes can be expressed as a combination of her effect on test scores and her effect on predicted non-cognitive ability (a factor that is a weighted average of several non-test score outcomes). In the second section, I use longitudinal survey data to form predicted non-cognitive ability, and I demonstrate the extent to which predicted non-cognitive ability in 8th grade predicts adult outcomes conditional on test scores. In the third section, I use administrative data to estimate 9th grade Algebra and English teacher effects on test-scores, and predicted non-cognitive ability. Using these estimates within the context of the model, I investigate how well test-score measures alone identify teachers that have large predicted effects on adult outcomes.

There are two distinct empirical challenges to credibly identifying teacher effects in high-school: The first challenge is that differences in outcomes across teachers may be due to selection of students to teachers; The second challenge is that *even with random assignment of students to teachers*, in a high-school setting, students in different tracks may be exposed to both different teachers and "track treatment effects" outside the classroom that influence classroom performance and confound teacher effects (e.g. students in the gifted track take Algebra with Mr. Smith and also take a time management class that has a direct effect on their performance on the

[2] Moreover, these studies are based on single cohorts of nationally representable samples that have too few student observations per school and teacher for credible identification of individual teacher effects.

Algebra I test). To address both these challenges, I follow Jackson (2012) and estimate models that condition on a student′s school-track (the unique combination of school, courses taken, and level of courses taken) so that all comparisons are made among students *at the same school and in the same academic track* — precluding any bias due to student selection to tracks or treatments that vary across tracks. In such models, variation comes from comparing the outcomes of students in the same track and school but who are exposed to different teachers either due to (a) changes in the teachers for a particular course and track over time, or (b) schools having multiple teachers for the same course and track in the same year. Because personnel changes within schools over time may be correlated with other changes within schools, I estimate models that also include school-by-year fixed effects. The remaining concern is that comparisons among students within the same track may be susceptible to selection bias. While most plausible stories of student selection involve selection to tracks rather than to teachers within tracks, I do present several empirical tests that suggest little to no selection bias.

Using the National Educational Longitudinal Study 1988 (NELS-88), a standard deviation increase in estimated non-cognitive ability in 8th grade (a weighted average of attendance, suspensions, grades, and grade progression) is associated with larger improvements in arrests, college-going, and wages than a standard deviation increase in test scores. Also non-cognitive ability is particularly important at the lower end of the earnings distribution. Using administrative data, 9th grade Algebra and English teachers have meaningful effects on test-scores, the same non-test score outcomes, and estimated non-cognitive ability. Additional calculations suggest that teacher effects on college going and wages may be as much as three times larger than that predicted based on test scores alone. As such, more than half of teachers who would improve long run outcomes *may* not be identified using test scores alone.

This is the first paper to demonstrate that teachers have meaningful effects on objective non-cognitive outcomes that are strongly associated with adult outcomes and are not correlated with test scores. This finding has important policy implications regarding measuring teacher quality, and suggests that one might worry that test-based accountability may induce teachers to divert effort away from improving students′ non-cognitive skills in order to improve test scores. Also, the finding that teachers have important effects on ability unmeasured by test scores offers a *potential* explanation for the empirical regularity that interventions that have test score effects that "fade out" over time can have lasting effects on adult outcomes (Chetty, Friedman, Hilger, Saez, Schanzenbach, & Yagan, 2011; Cascio & Staiger, 2012) .

The remainder of this paper is organized as follows: Section II presents the theoretical framework. Section III presents evidence that both test score and non-test score outcomes in 8th grade predict important adult outcomes. Sections IV and V present the data and empirical strategy, respectively. Sections VI and VII present the results and Section VIII concludes.

II Theoretical Framework

<u>*A Model of Multidimensional Teacher and Student Ability*</u>

Following Heckman, Stixrud, & Urzua (2006), I present a latent two-factor model where all student outcomes are a linear combination of students' cognitive and non-cognitive ability. Teachers vary in their ability to increase student cognitive and non-cognitive ability. Unlike a model with uni-dimensional student ability where a teacher who improves one outcome (such as test scores) necessarily improves *all* student outcomes (such as crime and adult wages), a model with multidimensional student ability has predictions that differ in important ways.

<u>Student ability:</u> Student ability is two-dimensional. One dimension is cognitive skills (e.g. content knowledge and computation speed), and the other dimension is non-cognitive or socio-behavioral skills (e.g. agreeableness, motivation, and self-control). Each student i has a two-dimensional 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.

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

<u>Outcomes:</u> There are multiple outcomes y_z observed for each student i. Each outcome y_z is a particular linear function of the ability vector such that $y_{zij} = \alpha_{ij}'\beta_z = (\nu_i + \omega_j)'\beta_z + \varepsilon_{zi}$ where $\beta_z = (\beta_{c,z}, \beta_{n,z})$ and $E[\varepsilon_{zi} \mid \varepsilon_{z'i}, \alpha, \omega] = 0$. Vector β_z captures the fact that while some outcomes may depend on cognitive ability (such as test scores), others may depend on non-cognitive skills (such as attendance). The variable ε_{zi} captures the fact that outcomes are measured with random student-level error. By definition, the two ability types are uncorrelated so that

$\text{cov}(\alpha_{c,ij}, \alpha_{n,ij}) = 0$. Test scores y_1 are a function of only cognitive ability so that $y_{1ij} = \alpha_{c,ij}\gamma + \varepsilon_{1i}$. 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 the parameter vector β_z is the factor loadings for student outcome *z*. The two-factor model is illustrated in a path diagram in Figure 1.

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 the average teacher*). Teacher *j*'s effect for outcome z is therefore $\theta_{zj} = \omega_j' \beta_z$. Teacher *j*'s effect for long run outcome y_* is $\theta_{*j} = \omega_j' \beta_* = \beta_{c,*}\omega_{c,j} + \beta_{n,*}\omega_{n,j}$ where $\beta_{c,*}\beta_{n,*} \neq 0$. The long-run outcome is not observed and policy-makers wish to predict teacher effects for long-run outcome y_*.

Proposition 1: *A teacher's effect on the long-run outcome may be correlated with her effect on multiple short-run outcomes, even if her effects on these short run outcomes are not correlated with each other.*

Consider the case with two outcomes y_1 *and* y_2. Each outcome reflects only one dimension of ability so that $\theta_{1j} = \beta_{c,1}\omega_{c,j}$ and $\theta_{2j} = \beta_{n,2}\omega_{n,j}$ where $\beta_{c,1}\beta_{n,2} \neq 0$. The two dimensions of teacher ability are uncorrelated so $\text{cov}(\omega_{c,j}, \omega_{n,j}) = 0$. In this scenario, the covariance between teacher effects across all three outcomes are given by [1] through [3] below.

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

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

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

Proposition 2: *With multiple short run outcomes that reflect a mix of both ability types, one can uncover estimates of a students' two-dimensional ability vector.*

Because the outcomes are linear combinations of the underlying ability types, any two outcomes are correlated *iff* they share the same mix of ability types. As such, outcomes that are based mostly on cognitive skills will be highly correlated and these will be weakly correlated with outcomes that are based largely on non-cognitive skills. For simplicity, we can standardize the variance of each ability-type so that $\text{var}(\alpha_{c,ij}) = \text{var}(\alpha_{n,ij}) = 1$. Given that

$y_z = \beta_{c,z}\alpha_{c,ji} + \beta_{n,z}\alpha_{n,ji} + \varepsilon_{zi}$, and $\text{cov}(\alpha_{c,ij}, \alpha_{n,ij}) = 0$, it follows that all covariance across outcomes are due to commonality in the underlying ability types across these outcomes so that $cov(y_1, y_2) = \text{cov}(\beta_{c,1}\alpha_{c,j} + \beta_{n,1}\alpha_{n,j} + \varepsilon_{1i}, \beta_{c,2}\alpha_{c,j} + \beta_{n,2}\alpha_{n,j} + \varepsilon_{2i}) = \beta_{c,1}\beta_{c,2} + \beta_{n,1}\beta_{n,2}$. As such, the covariance of the outcomes is given by [4], where E is the diagonal variance of the error terms.

$$\text{var}\begin{bmatrix} y^1 \\ \\ \\ y^z \end{bmatrix} = \Sigma = \begin{bmatrix} \beta_{c,1}\beta_{c,1} + \beta_{n,1}\beta_{n,1} & \cdots & \cdots & \beta_{c,1}\beta_{c,z} + \beta_{n,1}\beta_{n,z} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{c,1}\beta_{c,z} + \beta_{n,1}\beta_{n,z} & \cdots & \cdots & \beta_{c,z}\beta_{c,z} + \beta_{n,z}\beta_{n,z} \end{bmatrix} + E = \beta\beta' + E \quad [4]$$

It also follows that $cov(y_z, \alpha_c) = \text{cov}(\beta_{c,z}\alpha_{c,j} + \beta_{n,z}\alpha_{n,j} + \varepsilon_{zi}, \alpha_{c,j}) = \beta_{c,z}$ and similarly $cov(y_z, \alpha_n) = \text{cov}(\beta_{c,z}\alpha_{c,j} + \beta_{n,z}\alpha_{n,j} + \varepsilon_{zi}, \alpha_{n,j}) = \beta_{n,z}$. Accordingly, $\begin{bmatrix} Y \\ \alpha \end{bmatrix}$ is distributed with mean $\begin{bmatrix} \mu \\ 0 \end{bmatrix}$ and variance $\begin{bmatrix} \beta\beta' + E & \beta \\ \beta' & I \end{bmatrix}$. From Johnson and Wichern (2007), the conditional expectation of the underlying factor vector α is given by [5] below.[3]

$$E(\alpha \mid Y) = \beta'(\beta\beta' + E)^{-1}(Y - \mu) \quad [5]$$

As such, the best linear unbiased predictor of each underlying factor is a linear combination of each of the outcomes y such that $\hat{\alpha}_1 = m_1 Y$ and $\hat{\alpha}_2 = m_2 Y$, where m_1 and m_2 are defined from [5]. Put differently, one can predict student non-cognitive ability with a weighted average of their short run non-test-score outcomes.

Proposition 3: *One can express a teacher's overall effect on the long-run outcome as a linear combination of her effect on test scores and her effect on the predicted non-cognitive factor.*

By assumption, $y_{*ij} = \beta_{c,*}\alpha_{c,ij} + \beta_{n,*}\alpha_{n,ij} + \varepsilon_{*i}$. Because test scores are a proxy for the cognitive factor, $y_{1ij} = \alpha_{c,ij}\gamma + \varepsilon_{1i}$. Accordingly, $y_{*ij} = (\beta_{c,*}/\gamma)y_{1ij} + \beta_{n,*}\alpha_{n,ij} - (\beta_{c,*}/\gamma)\varepsilon_{1i} + \varepsilon_{*i}$. Because $E(\hat{\alpha}_{n,ij}) = \alpha_{n,ij}$, the conditional expectation of long run outcome based on test scores and

[3] The general statement is that if $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is distributed with mean $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and variance $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and $|\Sigma_{22}| > 0$. Then the conditional expectation of X_1 given $X_2=x_2$ is $E(X_1 \mid x_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$. See Johnson and Wichern (2007) for a formal proof. A condition for identification is that the number of outcomes be equal to or greater than the number of factors. As such, the two factors can be only be identified with more than one outcome.

predicted non-cognitive ability is $\hat{y}_{*ij} = E[y_{*ij} \mid y_{1ij}, \hat{\alpha}_{n,ij}] = (\beta_{c,*} / \gamma) y_{1ij} + \beta_{n,*} \hat{\alpha}_{n,ij}$. It follows that a teacher's predicted effect on the long run outcome is just her effect on the predicted outcome, which is a weighted average of her effect on test scores and her effect on the predicted non-cognitive factor. Specifically, $\hat{\theta}_{*j} = (\beta_{c,*} / \gamma)\theta_{1j} + \beta_{n,*}\theta_{\hat{\alpha}_c j}$, where θ_{1j} and $\theta_{\hat{\alpha}_c j}$ are teacher effects on test scores and predicted non-cognitive ability, respectively.

Discussion of the model: The model highlights that while effects on test scores may be a good measure of how much a teacher improves cognitive skills, it may not reflect how much a teacher improves non-cognitive skills. The model further indicates that non-test score outcomes that are sensitive to non-cognitive skills might allow one to predict students' non-cognitive ability and thus identify teacher effects on non-cognitive skills. Accordingly, because long run outcomes reflect a combination of both cognitive and non-cognitive skills, if we know the relationship between cognitive skills, non-cognitive skills, and long-run outcomes and we have a mix of test score and non-test score outcomes, one may be able to identify teachers that are excellent at improving long run student outcomes even if (a) their effect on long run outcomes are not directly observed and (b) such teachers have small effects on test scores. Hereinafter, I aim to use the implications of the model to form predictions of student non-cognitive ability, estimate teacher effects on both test scores and non-cognitive ability, and then form estimates of a teacher's effect on long-run outcomes based on her effects on short run outcomes.

Section III *Test Scores, Non-Cognitive Ability, and Long-Run Outcomes.*

In this section I use data from the National Educational Longitudinal Survey of 1988 (NELS-88) to test the underlying assumptions and implications of the model empirically. Specifically, I show that (a) there are two underlying factors that explain most of the covariance between test score and non-test score outcomes in 8th grade that can be identified as cognitive and non-cognitive ability; (b) test scores are a good proxy for cognitive ability; (c) both higher test scores and higher non-cognitive ability are associated with better adult outcomes even though they are uncorrelated; (d) teachers effects on long run outcomes can be expressed as a combination of effects on test scores and the non-cognitive factor.

The NELS-88 is a nationally representative sample of respondents who were eighth-graders in 1988. These data contain information on short run outcomes (absences, suspensions, GPA, grade repetition, and math and reading test scores) and long run outcomes (being arrested or having a close friend who has been arrested, attending a post-secondary institution, and income in 1999 when most respondents were 25 years old). The top panel of Table 1 presents the correlations between the short run outcomes in these data.

The first notable pattern is that test score outcomes are relatively strongly correlated with each other (math scores and reading scores have correlation≈0.8), are moderately correlated with grade point average (correlation≈0.38 for math and correlation≈0.34 for reading), and are weakly correlated with other non-test-score outcomes (the average absolute value of the individual correlations is 0.098). This is suggestive of an underlying cognitive factor that is highly predictive of standardized test scores, is moderately related to grades, and is largely unrelated to socio-behavioral outcomes. The other pattern is that the non-test score outcomes are much more strongly correlated with each other (average correlation of 0.16) than with the test score outcomes (the average absolute value of the individual correlations is 0.098), and are moderately correlated with GPA (correlation of about 0.35). This suggests that there is a non-cognitive factor that explains the correlations between the non-test score outcomes, is an important determinant of grades, and is unrelated to test scores. The fact that GPA is correlated with both sets of variables is consistent with research (e.g. Howley, Kusimo, & Parrott, 2000; Brookhart, 1993) finding that most teachers base their grading on some combination of student product (exam scores, final reports etc.), student process (effort, class behavior, punctuality, etc.) and student progress — so that grades reflect a combination of cognitive and non-cognitive skill.

To formally assess the degree to which the variables can be classified as measuring cognitive and non-cognitive skills, I estimate the factor analytical model outlined in section II. From the model we know that $E(\alpha \mid Y) = \beta'(\beta\beta' + E)^{-1}(Y - \mu)$ and the covariance of the short run outcomes is $\beta\beta' + E$. With a sufficiently large sample, the sample covariance matrix S is a consistent estimate of $\beta\beta' + E$. If the error vector is normally distributed, one can estimate the factor loadings β by maximum likelihood.[4] Accordingly, as suggested in Johnson and Wichern (2007) I use $\hat{\alpha} = \hat{\beta}'\hat{S}^{-1}(Y - \overline{y})$ as my estimate of the underlying factor vector. If the model presented in Section II is reasonable, there should be two estimated factors that are identifiable as measuring cognitive and non-cognitive skills.

The first factor explains 90 percent of the covariance in the short run outcomes, the second explains 9 percent of the covariance— consistent with two underlying factors that explain covariation in student outcomes (Heckman, Stixrud, & Urzua 2006). The correlation between the first factor and math and reading scores is 0.95 and 0.95 respectively, the correlation with GPA is 0.4, and the correlation with other outcomes are all less than 0.2 in magnitude —consistent with this first factor being cognitive skills and test scores being a good proxy for cognitive skills. The second factor is strongly correlated with the other non-test score outcomes and weakly correlated with test scores. *Because I wish to capture the importance of changes in non-cognitive ability that is not predicted by test scores*, *I regress this factor on test scores and take the residuals as my measure of non-cognitive ability*.[5] I present the correlation between the second factor and the outcomes in Table 1. This second factor is uncorrelated with test scores (by construction), has a strong/moderate negative correlation with being absent and being suspended, and has a strong positive correlation with GPA and attending 9th grade on time. This factor explains much variability in socio-behavioral outcomes and is not correlated with test scores. It is reasonable to call this factor non-cognitive skill that is not detected by standardized tests.

While I am agnostic about what exact set of skills are captured by this factor, studies suggest that each of these non-test score outcomes is associated with the same personality traits. Psychologists typically classify non-cognitive traits in terms of five dimensions; Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. Low levels of agreeableness and high neuroticism are associated with higher school absences, higher externalizing behaviors, greater 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, &

---

[4] The likelihood of the data is $L(\mu,\Sigma) = (2\pi)^{\frac{-np}{2}} |\Sigma|^{\frac{-n}{2}} e^{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\overline{x})(x_j-\overline{x})'+n(\overline{x}-\mu)(\overline{x}-\mu)'\right)\right]}$, where p is the number of short run outcomes. Because the covariance matrix $\Sigma = \beta\beta' + E$ is a function of the factor loading vector β, the maximum likelihood estimate $\hat{\beta}$ of β is the one that maximizes this likelihood (Johnson and Wichern 2007).

[5] Note that the correlation between the unadjusted non-cognitive factor and test scores was 0.18.

Goodman, 2007). High conscientiousness is associated with fewer absences, fewer externalizing behaviors, higher grades, more on-time grade progression, and higher educational attainment (Duckworth, Peterson, Matthews, & Kelly, 2007). This suggests that the factor explaining covariation between absences, grade progression, suspensions, and grades reflects a skill-set that is associated with high conscientiousness, high agreeableness, and low neuroticism. Consistent with this factor being unrelated to test scores, cognitive ability is associated with high openness, but largely unrelated to the other traits (Furnham, Monsen, & Ahmetoglu, 2009).

In Table 2, I show that the two uncorrelated measures of cognitive and non-cognitive ability independently predict long run outcomes. I regress various adult outcomes on these two ability measures. For ease of interpretation, both test scores and the non-cognitive factor have been standardized to be mean zero unit variance. To remove correlation between these outcomes and adult outcomes due to differences in socioeconomic status or demographic differences *all models include controls for household income, whether English is the primary language at home, whether the student lives with their mother or father or both, the mothers and father highest level of education, family size, student race and student gender.* Column 3 shows that while test scores have little relation to whether a student is arrested or has had a close friend who was arrested, a one standard deviation increase in the non-cognitive factor is associated with a 5 percentage point reduction (this is a 25% percent reduction relative to the sample mean). Columns 1 and 2 present results for the individual non-test score outcomes. Column 6 shows the effect on college attendance based on a linear probability model. The coefficient indicates that a one standard deviation increase in test scores is associated with a 7.6 percentage point increase in college going, while a one standard deviation increase in the non-cognitive factor is associated with a 10.7 percentage point increase in college going (the baseline level is 0.79). Column 9 shows the effect on the natural log of earnings based on unconditional quantile regression (Firpo, Fortin, & Lemieux, 2009) evaluated at the median of the outcome distribution.[6] The coefficient indicates that a one standard deviation increase in test scores is associated with 5.55 percent higher earnings, while a one standard deviation increase in the non-cognitive factor is associated with 5.97 percent higher earnings. Note that using the standardized cognitive factor instead of standardized test scores yields a coefficient of 0.048 on the cognitive factor. *Because the*

[6] Note that I only look at income, conditional on any income data. Note that roughly 90 percent of observations have income data so that quantile regression results are similar for any reasonable treatment of missing data.

cognitive factor and test scores are highly correlated and interchangeable in predicting long run outcomes, and it is standard practice to estimate teacher effects on test scores, I use test scores as my measure of cognitive ability for the remainder of the paper.

In recent findings, Lindqvist & Vestman (2011) and Heckman, Stixrud, & Urzua (2006) find that non-cognitive ability is particularly important at the lower end of the earnings distribution. To ensure that my measure of non-cognitive ability is valid, it is important to show similar relationships in these data. To test for this, I estimate quantile regressions to obtain the marginal effect on log wages at different points in the earnings distribution and present the estimated coefficients in Figure 2. The results are strikingly similar to Lindqvist & Vestman, (2011). Specifically, point estimates suggest that cognitive ability is similarly important over all parts of the earnings distribution. In contrast, non-cognitive ability is very important at the lower end of the earnings distribution and becomes less important at higher ends of the distribution. This is consistent with the effects on arrests and echoes findings in the literature, thereby suggesting that this factor is a good measure of non-cognitive ability.

The results indicate that student outcomes are indeed a reflection of *both* cognitive and non-cognitive ability, and in many cases, the effect of non-cognitive ability is larger than that of cognitive ability. Accordingly, teachers who improve non-cognitive ability may have important effects on adult outcomes even if they have no effects on test scores. In light of these relationships, the remainder of the paper aims to estimate the extent to which 9th grade teachers improve both test scores and predicted non-cognitive ability. Using these estimates, I then aim to uncover the effect of teachers on long run outcomes through both dimensions of ability.

IV Data:

To estimate the effect of teachers on student outcomes, this paper uses data on all public middle- and high-school students in North Carolina from 2005 to 2010 from the North Carolina Education Research Data Center. The student data include demographics, transcript data on all courses taken, middle-school achievement data, 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.[7] I limit the analysis to students who took either the Algebra I or English I

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

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 that the resulting sample is representative of 9th graders as a whole. To avoid endogeneity bias that would result from teachers having an effect on 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 3.

These data cover 348547 9th grade students in 619 secondary schools in classes with 4296 English I teachers, and 3527 Algebra I teachers. While roughly half of the students are male, about 58 percent are white, 29 percent are black, 7.5 percent are Hispanic, 2 percent are Asian, and the remaining one percent is Native American, mixed race, or other. About 7.5 percent of students have the highest parental education level (i.e. the highest level of education of the student's two parents) below high-school, 40 percent with a high school degree, about 15 percent with a junior college or trade school degree, 20 percent with a four year college degree or greater, and 6.4 percent with an advanced degree (about 10 percent of students are missing data on parental education). The achievement variables have been standardized to be mean zero with unit variance for each cohort and test. Incoming 7th and 8th grade test scores in the final 9th grade sample are approximately 8 percent of a standard deviation higher than that of the average in 7th or 8th grade. This is because the sample of 9th grade students is less likely to have repeated a grade and to have dropped out of the schooling system.

Looking to the outcomes, the average number of absent days is 3 and the average of the log of days absent is 0.585. About 85 percent of students were in 10th grade the following year, and about 5.6 percent of all 9^{th} graders had an out of school suspension. The lower panel of Table 1 shows the correlations of these outcomes in the NCERDC data. These variables have a similar correlational structure to the NELS-88 data such that students with better test scores do not necessarily have better non-test score outcomes. Specifically, while test scores are strongly correlated (corr = 0.616), they are moderately correlated with GPA (corr ≈ 0.55). Test scores have a moderate correlation with being in 10^{th} grade on time (corr ≈ 0.32), and weak correlations with absences, and being suspended. To obtain a measure of non-cognitive skills that improves adult outcomes *but is not captured by test scores*, I recreated the non-cognitive factor in the NCERDC data using the factor weights derived from the NELS-88 data and removing any correlation with test scores. As such, *improvements in this measure capture improvement in real* *skills that are not detected by improvements in test scores*. The correlations between the non-cognitive factor and the outcomes are similar to those in the NELS-88 data.[8]

Measuring Tracks

Even though schools may not have explicit labels for tracks, most do 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 (2012) and Harris & Anderson (2012), it is not just the course that matters but also the level at which students take a course. Indeed, even among students taking Algebra I or English I courses, there are three different levels of instruction (advanced, regular, and basic). As such, I exploit the richness of the data and take as my measure of a school-track, 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 who take the same academic courses, and take the same level of English I, and the same level of Algebra I, all at the same school are in the same school-track.*[10] Defining tracks flexibly at the school-by-course-group-by-course-level level allows for different schools to have different selection models and treatments for each track. Because many students pursue the same course of study, only 3.7 percent of all student observations 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.

V Empirical Strategy

This section outlines the strategy to estimate teacher effects on test score outcomes, non-test-score outcomes, and predicted non-cognitive ability. The empirical approach is to model student outcomes as a function of lagged student achievement and student covariates, with the additional inclusion of controls for student selection to tracks and any treatments that are specific

[8] Note: The correlation between the non-cognitive factor as predicted using the correlational structure in the NCERDC data and the non-cognitive factor as predicted using the correlational structure in the NELS data is 0.98. Even though the results are similar, because I aim to use the same factor that is demonstrated to improve adult outcomes in the NELS-88 data, I create the factor using the weights derived from the correlational structure in the NELS-88 data.

[9] While there are hundreds of courses that students can take (including special topics, and reading groups) there 10 core academic courses that constitute about 66 percent of all courses taken. they are listed in Appendix Table A1.

[10] Students who take the same set of courses at another school are in different school-tracks, students who are in the same school but took either a different number of courses or at least one different course are in different school-tracks, and students at the same school who took the same courses but took Algebra I or English I at different levels would are in different school-tracks.

to tracks that might affect student outcomes directly. In such models I remove the influence of track-level treatments and selection to tracks on estimated teacher effects comparing student outcomes across teachers within groups of students *in the same track at the same school*. Specifically, I model the outcomes Y_{icjgys} of student i in class c with teacher j in school-track g, at school s, in year y with [6] below (note: most teachers are observed in multiple classes).

$$Y_{icjgys} = A_{iy-1}\delta + X_i\beta + I_{ji}\cdot\theta_j + I_{gi}\,\theta_{\mathrm{g}} + I_{sy}\,\theta_{\mathrm{sy}} + \phi_{\mathrm{c}} + \varepsilon_{icjgys} \qquad [6]$$

A_{iy-1} is a matrix of incoming achievement of student i (8^{th} grade and 7^{th} 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_{gi}, an indicator variable equal to 1 if student i is in school-track g and *0* otherwise so that θ_g is a time-invariant fixed effect for school-track g,[11] I_{sy} is an indicator variable denoting whether the student observation 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 mean zero random error term. In these models, 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, school-by-year level shocks and school-by-track effects. For test score outcomes, this model is just a standard value-added model with covariate adjustments.

Because the main 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 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 the identities of Algebra I and English I teachers over time due to staffing changes 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

[11] Note: In expectation, the coefficient on the school-track indicator variable reflects a combination of *both* the unobserved treatment specific and selection to school-track g.

outcomes of students in track B with teacher 2 in 2000 with those of students in tracks B 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 in [6] (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. In section VI, I test for this explicitly and find little evidence of 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-tracks. In section VI, I show that the findings are not driven by student selection within school-tracks.

To provide a sense of how much variation there is within school-tracks during the same year versus how much variation there is within school-tracks across years, 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, so that for more than half the data the variation is based on comparing single teachers across cohorts within the same school-track. It follows that more than one-third of the variation is within tracks-year cells. In section V.4 I show that results using all the variation are similar to those using only variation across cohorts.

Illustrating the importance of accounting for tracks

Because including school-track effects is nonstandard, it is important to illustrate the importance of conditioning on tracks. I do this by showing how conditioning on school-tracks affects the relationship between teacher experience and test scores. To do this, I regress English I and Algebra I scores on 8th and 7th grade math and reading scores and indicator variables for each year of teacher experience. I estimate models with school fixed effects, and then with track-by-school fixed effects. I plot the estimated coefficients on the years of experience indicator variables in Figure 3. Across all models, students perform better with Algebra teachers who have more years of experience (top panel). Models with track-by-school effects are about 25 percent smaller than those without. The results for English are even more dramatic such that models with track-by-school effects are 66 percent smaller than those without. This suggests that either (a) more experienced teachers select into tracks with unobservably higher achievement students, or (b) tracks with more experienced teachers provide other supports (outside the teachers own classroom) that directly improve outcomes— indicating that failing to account for school-tracks may lead to much bias. To assuage concerns that the reduction is due to a lack of variation within school-tracks, I compute the fraction of the variance in experience overall and within school-tracks. For both subjects, about 60 percent of the variation in teacher experience occurs within school-tracks. Accordingly, a lack of variability within tracks *cannot* explain the reduced effects— underscoring the importance of accounting for school-tracks.

Estimating the Variance of Teacher Effects

It is known that the variance of the estimated teacher effects $\hat{\theta}$ from [6] will overstate the variance of true teacher quality because of sampling variation and classroom level shocks. As such, I follow Kane & Staiger (2008), Jackson (forthcoming) and Chetty, Friedman, & Rockoff (2011) 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:

<u>*Step 1:*</u> I estimate equation [7] below.

$$Y_{icjgys} = A_{iy-1}\delta + X_i\beta + I_{gi}\,\theta_{\mathrm{g}} + I_{sy}\,\theta_{\mathrm{sy}} + \phi_{\mathrm{c}} + \theta_j + \varepsilon_{icjgys} \qquad [7]$$

There are no teacher 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 [7] 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.

<u>*Step 2:*</u> I link every classroom-level mean residual and pair it with a random different classroom level mean residual for the same teacher and compute the covariance of these mean residuals. That is, I compute $Cov(\bar{e}_c^{*}, \bar{e}^{*}_{c'} \mid J = j)$. To ensure that the estimate is not driven by any particular random pairing of classrooms within teachers, I replicate this calculation 100 times and take the median of the estimated covariance as my parameter estimate. If the classroom errors ϕ_{c} are uncorrelated with each other (note: the model includes school-by-year fixed effects) and are 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). That is, $Cov(\bar{e}_c^*, \bar{e}*_{c'} \mid J=j) = \text{cov}(\theta_j, \theta_j) = \text{var}(\theta_j) \longrightarrow \sigma^2_{\theta_j}$.[12]

As discussed in Jackson (2012) because the variance of the true teacher effects can be estimated by a sample covariance, one can compute confidence intervals for the sample covariance. I use the empirical distribution of 100 randomly computed "placebo" covariances (i.e. sample covariance across classrooms for *different* teachers) to form an estimate of the standard deviation of the sample covariance across classrooms for the same teacher. I then use this "bootstrapped" standard deviation of the covariance for normal-distribution-based hypothesis testing. Because most studies report the standard deviation of teacher effects, I report the square root of the sample covariance and the square root of the confidence bounds.

VI Main Results

Table 5 presents the estimated covariance across classrooms for the same teachers under three different models. Because standard deviations are positive by definition, when the sample covariance is negative, I report the standard deviation to be zero (*note that none of the negative covariance estimates is statistically significantly different from zero at the 5 percent level*). I present naïve models that include school fixed effects and attempt to account for tracking with peer characteristics (mean peer 8th and 7th grade math and reading scores in addition to mean peer demographics). I then present models that account for tracking with school-track fixed effects. And finally I estimate 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.

In models that include school effects and peer characteristics, the estimated standard deviation of algebra teacher effects on algebra test scores (left panel) is 0.12. This is similar to estimates found in other studies of high school teacher that do not account for tracking explicitly. In this naïve model, algebra teachers have statistically significant effects on all outcomes except for English scores and the standard deviation of Algebra teacher effects on non-cognitive ability is 0.16. Models for English teachers (right panel) indicate that the standard deviation of English teacher effects on English scores is 0.049. In this naïve model, English teachers have statistically significant effects on all non-test score outcomes and the standard deviation of English teacher

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

effects on non-cognitive ability is 0.135.

The middle panel presents results that include track-by-school fixed effects instead of peer characteristics to account for tracking. In such models, the estimated effects fall by about 20 percent relative to only including peer level characteristics. Adding the additional controls for school-by-year effects reduces the effects by a further 30 to 50 percent— suggesting that there may be bias associated with omitting both school-track effects and not accounting for school-by year shocks. In the preferred model (lower panel), the standard deviation of the algebra teacher effect on algebra test scores is 0.066; an estimate about half the size as that obtained in the naïve model that does not account for tracking or school specific shocks. The estimated teacher effects are statistically significantly different from zero for some non-cognitive outcomes such that the standard deviation of teacher effects on GPA is 0.045, on enrolling in 10th grade is 0.025, and the standard deviation of the effects on the non-cognitive factor is 0.049. Looking to English I teachers, in the preferred model (lower panel), the standard deviation of the English teacher effect on English test scores is 0.03. While there is no effect of English teachers on algebra scores, the estimated teacher effects are statistically significantly different from zero for all the non-cognitive outcomes such that the standard deviation of teacher effects on being suspended is 0.014, the effect on log of absences is 0.036, that on GPA is 0.027, that on enrolling in 10th grade is 0.024, and the standard deviation of the effects on the non-cognitive factor is 0.049 with a 95 percent confidence interval of [0.025, 0.079].

To put the non-test score estimates into perspective, having an algebra or English teacher at the 85th percentile of GPA quality versus one at the 15th percentile would be associated with 0.09 and 0.054 higher GPA, respectively. For both subjects, an Algebra or English teacher at the 85 percentile of on-time grade progression quality versus one at the 15th percentile would be associated with being 5 percentage points more likely to enroll in 10th grade on time (an effect size of 0.14σ). Given that not enrolling in 10th grade is a very strong predictor of dropout, this suggests significant teacher effects on dropout. These findings are consistent with, but about half the size of those dropout effects found in Koedel (2008). The English teacher effects are such that having an English teacher at the 85th percentile of absences and suspensions quality versus one at the 15th percentile would be associated with being 2.8 percentage points less likely to be suspended (an effect size of 0.12σ), and have 7.4 percent fewer days absent. Overall, for both subjects, having a teacher at the 85th percentile of improving non-cognitive ability versus one at the 15th percentile would be associated with 0.098σ higher non-cognitive ability; effect sizes that are similar to those on test scores. If non-cognitive ability is as important in determining long run outcomes as cognitive ability (as suggested in Table 2), and there is no correlation between effects on both kinds of ability, then test-score-based measures of teacher quality may drastically understate the importance of teachers for long run outcomes. I investigate this in section VII.

Tests for Bias due to Selection

While the inclusion of student covariates and student achievement both one and two years before should account for bias due to dynamic tracking (Rothstein, 2010), school-by-track effects should account for bias due to omitted school-track treatment or selection to school-tracks (Jackson, 2012), and school-by-year effects should remove bias due to school specific time shock that may be confounded with teacher effects, readers may still worry that the estimated results are biased by student selection to teachers. I follow Chetty, Friedman, & Rockoff (2011) and present two tests for selection for the two main outcomes (test scores and non-cognitive ability). In the first test, I exploit the statistical fact that any selection within school-track-years will be eliminated by aggregating the treatment to the school-track-year level (leaving only variation across years/cohorts within school-tracks). If the effects are driven by students selecting to individual teachers within school-track years, then the marginal effects of estimated teacher quality (estimated using all *other* years of data) using variation *across* cohorts but within school-track should be equal to zero. Conversely, if the effects using only variation across cohorts in the same school-track are similar to those obtained using all the variation, it would suggest little selection bias. To test this, I estimate equations [8] and [9] on the data where $\hat{\theta}_j$ is the estimated (out of sample)[13] effect of teacher j, $\bar{\hat{\theta}}_{sgy}$ is the mean estimated teacher effect in school-track g in year y. As before, θ_{g} is a school-track effect and θ_{sy} is a school year effect.

$$Y_{ijcy} = A_{iy-1}\delta + \psi_1\hat{\theta}_j + X_{iy}\beta + X^*_{jy}\pi + I_{gi}\,\theta_{\mathrm{g}} + \theta_{\mathrm{sy}} + \varepsilon^*_{ijcy} \qquad [8]$$

$$Y_{ijcy} = A_{iy-1}\delta + \psi_2\bar{\hat{\theta}}_{sgy} + X_{iy}\beta + X^*_{jy}\pi + I_{gi}\,\theta_{\mathrm{g}} + \theta_{\mathrm{sy}} + \varepsilon^*_{ijcy} \qquad [9]$$

In [8] because teacher quality is defined at the student level, variation comes from both comparing students in the same school-track across different cohorts and comparing students in

[13] 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.

the same school-track in the same year who have different teachers. In contrast, by defining the treatment at the school-track-year level in [9], one is no longer comparing students within the same school-track-year, but only comparing students in the same school-track across different cohorts where selection is unlikely. Conditional on school-track fixed effects, all the variation in this aggregate teacher quality measure in [9] occurs due to changes in the identities of teachers in the track over time. That is, equation [9] tests for whether cohorts on average perform better/worse with the arrival/departure of a teacher who has been found to be effective with other students in other years. If there is no sorting in unobserved dimensions within tracks, ψ_1 from [8] should be roughly equal to ψ_2 from [9]. However, if all of the estimated effects are driven by sorting within school-track-years, then the out of sample estimates will be spurious and there should be no effect on average associated with changes in school-track-cohort mean teacher value-added in the track so that ψ_2 from [9] will be equal to 0. Because the estimated teacher effects are estimated with noise, the coefficients ψ_1 and ψ_2 will be less than 1.

The results of this test are presented in Table 6. For Algebra teacher effects on Algebra scores, using all the variation (column 2) yields a point estimate of 0.283, and using changes in aggregate track level mean teacher quality (using only variation across years) yields a point estimate of 0.279. The null hypothesis that the Algebra teacher effect on algebra scores is driven by selection within cohort is rejected at the 1 percent level. Recall that this model includes school-by-year fixed effects so the results are not confounded by school level shocks that may be associated with changing identities of teachers. The point estimates indicate that teachers who improve test scores by 1σ out of sample increase test scores by only 0.28σ. This reflects the fact that teacher effects are measured with considerable error so that roughly one third of the teacher effect is signal (this is similar to other studies). Looking to Algebra teacher effects on non-cognitive ability (column 4), using all the variation yields a point estimate of 0.0603, and using changes in track level mean teacher quality (only variation across years) yields a point estimate of 0.0628. The effects are similar and the null hypothesis that the Algebra teacher effect on non-cognitive ability is driven by selection within cohort is rejected at the 5 percent level.

The results for English teacher tell a similar story. English teacher effects on English scores, using all the variation (column 6) yields a point estimate of 0.214, and using changes in aggregate track level mean teacher quality yields a point estimate of 0.230. The null hypothesis that English teacher effect on English scores is driven by selection within cohort is rejected at the

1 percent level. Looking to the English teacher effect on non-cognitive ability, using all the variation (column 8) yields a point estimate of 0.0468, and using changes in track level mean teacher quality yields a point estimate of 0.0782. The null hypothesis that effects on non-cognitive ability are driven by selection within cohorts is rejected at the 1 percent level.

The second test for selection is to see if estimated teacher effects are correlated with observable characteristics that predict the outcomes. To test for this, I follow Chetty, Friedman, & Rockoff (2011) and Jackson (2012) and regress the outcomes on all demographic and student achievement covariates and create a predicted outcome. I then estimate equations [8] and [9] on predicted outcomes rather than the actual outcomes (with no controls). As one can see in columns 1, 3, 5, and 7, none of the models yield relationships that are statistically significant at the 5 percent level, and many of the small negative point estimates indicate that any bias goes in the opposite direction of the estimated effects.

VII *Relationship between Teacher Effects across Outcomes*

Having established that teachers have real casual effects on test scores and predicted non-cognitive ability (that is uncorrelated with test scores), in this section I aim to document the extent to which test score based value-added understates the importance of teachers on long-run outcomes. To gain a sense of whether teachers who improve cognitive ability also improve other outcomes, I regress the estimated teacher effects for all the outcomes on the Algebra and English test score effects and report the R^2 in Table 7. The reported R^2 measure the fraction of teacher effects on each outcome that can be explained by (or is associated with) effects on test scores. The top row presents effects for Algebra teachers. For all outcomes, teachers with higher test score value-added are associated with better non-test score outcomes, but the relationships are weak. Algebra teacher effects on algebra test scores explain 0.77 percent of the variance in estimated teacher effects on suspensions, 2.5 percent of the estimated effect on absences, 10 percent of the effect on GPA, and 5 percent of the effect on on-time 10th grade enrollment (top panel top row). This indicates that while teachers who raise test score may also be associated with better non-test-score outcomes, the lion's share of effects on non-test score outcomes is unrelated to effects on test scores. Given that the non-cognitive factor is a weighted average of these non-test score outcomes and is unrelated to test scores by construction, it is not surprising that a teacher's effects on Algebra test scores explains only 0.91 percent of the variance of estimated teacher effects on non-cognitive ability. In contrast to test score value-added, effects on non-cognitive ability explain a sizable share of the estimated effects on the non-test score outcomes. Specifically, algebra teacher effects on non-cognitive ability explain 16 percent of the variance in estimated teacher effects on suspensions, 37 percent of the estimated effect on absences, 56 percent of the effect on GPA, and 26 percent of the effect on on-time 10th grade enrollment (top panel second row). As expected, teacher effects on non-cognitive ability explain only 0.91 percent of the variance in estimated teacher effects on Algebra test scores.

Results for English teachers follow a similar pattern. 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 only 0.89, 0.73, 4.9, and 6.6 percent of the variance in estimated teacher effects on suspensions, absences, GPA, and on-time 10th grade enrollment, respectively (lower panel top row). As expected, a teacher's effects on English test scores explains 1.86 percent of the variance in estimated teacher effects on non-cognitive ability. In contrast to test scores, English teacher effects on non-cognitive ability explain a sizable share of the estimated effects on the non-test score outcomes. Specifically, English teacher effects on non-cognitive ability explain 26.3 percent of the variance in estimated teacher effects on suspensions, 40.13 percent of estimated effects on absences, 56 percent of effects on GPA, and 32.8 percent of effects on on-time 10th grade enrollment (lower panel second row).

In sum, the results indicate that a teacher's effect on test scores and other non-cognitive outcomes are largely orthogonal, such that teachers who tend to improve test scores are no more or less likely to improve non-test score outcomes. To show this visually, Figure 4 presents a scatterplot of the estimated test score effects and effects on non-cognitive skills. It is clear that a teacher's effect on non-cognitive skills is essentially missed by her effect on test scores. This implies that roughly half of teachers classified as above average at improving test score will be below average at improving non-cognitive ability and roughly 25 percent of teachers in the top 25 percent of improving test scores will be in the bottom 25 percent at improving non-cognitive ability. Because unexplained variability in outcomes associated with individual teachers is not just noise, but is systematically associated with their ability to improve unmeasured non-cognitive skills, classifying teachers based on their test score value-added will likely lead to large shares of excellent teachers being deemed poor and vice versa. This also implies that if effects on non-cognitive ability are associated with meaningful improvements in adult outcomes, there could be considerable gains associated with measuring teacher effects on both test score and non-test score outcomes over using test score measures alone. Another implication is that if teachers face tradeoffs such that they must expend less effort improving non-cognitive ability in order to improve cognitive ability, regimes that increase the relative external rewards for test scores (such as paying teacher for test score performance or test based accountability) may undermine the creation of students' non-cognitive skills (Holmstrom & Milgrom, 1991). In light of the large estimated benefits to higher non-cognitive skills (particularly for students at the lower end of the earnings distribution) in Table 2, this may be cause for concern. Finally, the results suggest that the importance of teachers on student long-run outcomes may be much larger than that suggested by comparing the adult outcomes of students with high versus low value-added teachers.

How much do Test Score Measures Understate the Importance of Teachers?

The results thus far have shown that (a) variation in non-cognitive outcomes may be more determinative of adult outcomes than test scores, (b) teachers have effect sizes on non-cognitive ability that are similar to their effects on test scores, and (c) teacher effects on test scores and their effects on non-cognitive skills are largely unrelated. Taken together, this implies that test score based measures of teacher quality will vastly understate the true importance of teachers for long run outcomes and many of the most effective teachers will not be identified based on test score based measures. To see this point, consider this thought exercise. If cognitive ability and non-cognitive ability are uncorrelated, are equally important in determining wages, and teachers had the exact same effects sizes on cognitive and non-cognitive ability, then teacher effects on student wages as measured by test scores would reflect roughly 50 percent of a teachers' overall effect. If teacher effects were larger on non-cognitive ability than for cognitive outcomes (as is the case for college going) test scores effects would explain less than 50 percent of the overall effect on wages. The exact extent to which test score value-added explains variability in a teachers' overall effect depends on the relationship between the importance of cognitive ability (in this case test scores) and non-cognitive ability in determining adult outcomes and the exact relationship between teacher effects on both dimensions of ability in the data.

As discussed in Section II, one can express a teacher's overall predicted effect on long-run outcomes as a linear combination of her effect on test scores and her effect on predicted non-cognitive ability. Accordingly, we can predict a teacher's effect on college-going or wages based on her effect on test scores and predicted non-cognitive ability *as long as we know the marginal* *effect of test scores and non-cognitive ability on wages*. While there is no way to know for certain, the results in Table 2 present some guidance. Specifically, column 6 of Table 2 indicates that the marginal effect of increasing test scores (cognitive ability) by 1sd is to increase college going by 7.67 percentage points, and the marginal effect of increasing non-cognitive ability by 1sd is to increase college-going by 10.7 percentage points. One could compute a teacher's predicted effect on college-going based on both her contribution to cognitive skills and non-cognitive skills as $\hat{\theta}_j^* = (0.0767)\hat{\theta}_j^{scores} + (0.107)\hat{\theta}_j^{non-cognitive}$. The R-squared of a regression of $\hat{\theta}_j^*$ on $\hat{\theta}_j^{scores}$ will provide a measure of how much of the predicted effect on college-going can be explained by a teacher's effect on test scores. Similarly, using estimates from column 9 of Table 2 one could compute a teacher's predicted effect on log wages based on both her contribution to cognitive skills and non-cognitive skills as $\hat{\theta}_j^* = (0.0555)\hat{\theta}_j^{scores} + (0.0597)\hat{\theta}_j^{non-cognitive}$. The R-squared of a regression of $\hat{\theta}_j^*$ on $\hat{\theta}_j^{scores}$ will provide a measure of how much of the predicted effect on log wages can be explained by a teacher's effect on test scores.

The problem with simply using the estimated relationships from the NELS-88 data is that the relationships are based on correlations and the variables are measured with error, so one cannot be certain that these represent causal relationships. As such, I construct teacher's predicted effect on college-going and wages under different assumptions about the relative weight for non-cognitive ability. That is, when the relative non-cognitive weight for college going is *r* then, $\hat{\theta}_j^{college} = (0.0767)\cdot(\hat{\theta}_j^{scores} + r\cdot\hat{\theta}_j^{non-cognitive})$. Similarly, when the relative non-cognitive weight for log wages going is *r* then $\hat{\theta}_j^{\log wages} = (0.0555)\cdot(\hat{\theta}_j^{scores} + r\cdot\hat{\theta}_j^{non-cognitive})$. The relative weight for college-going from the empirical estimates is 0.107/0.0767=1.39 and that for log wages is 0.0597/0.0555=1.075. This is consistent with several studies finding that non-cognitive abilities are at least as important as cognitive skills (e.g. Bowles, Gintis, & Osborne, 2001; Jencks, 1979; Heckman, Stixrud, & Urzua, 2006).

To show how much of the variability in a teacher's predicted effect on college going and wages is explained or detected by her effect on test scores, Figure 5 plots the R^2 of a regression of $\hat{\theta}_j^*$ on $\hat{\theta}_j^{scores}$ for different weights. *Note that because the R^2 is invariant to the unit of measurement, the pattern in Figure 5 will hold for any outcome and any marginal effect of test* *scores on college-going or wages.* For Algebra teachers, by construction, when the relative weight is 0, the estimated teacher effect on test scores predicts all of the predicted teacher effect on wages. With a weight of 50 (such that non-cognitive skills are half as important as cognitive skills in determining the outcome) Algebra test score effects explain about 70 percent of the full effect. At values closer to the consensus value for wages of 1, Algebra test score effects explain about 30 percent of the full effect. For the empirical weight for college going of 1.39, Algebra test score effects explain only about 18 percent of the full effect. Because the relative effects are smaller on English test scores than for algebra test scores, for English teachers the results are even more dramatic. With a weight of 0.50, English test score effects explain about 55 percent of the full effect. At values closer to the consensus value for wages of 1, English test score effects explain only about 20 percent of the teacher's full effect on wages. For the empirical weight for college going of 1.39, English test score effects explain only about 10 percent of the full teacher effect. This drop stems from the fact that as the weight on non-cognitive ability increases there is more variability in predicted college-going and wage effect that is uncorrelated with test score effects so that the variability explained by effects on test scores approaches zero.

These results suggest that a teacher's effect on college-going and earnings as measured based on her effect on test scores alone may reflect between 10 and 30 percent of her overall effect. In the context of recent findings that having a teacher at the 85th percentile of the test score value-added distribution versus a median teacher is associated with students being 0.5 percentage points more likely to attend college and earning \$186 more per year (Chetty et. al. 2011), calculations suggest that having a teacher at the 85th percentile of the quality distribution for college-going and wages versus a median teacher may be associated with being between 1 and 5 percentage points more likely to attend college and earning between \$600 and \$1000 more per year. While the estimated range of values is large, they illustrate that estimated teacher effects based on test scores likely drastically understate the overall importance of teachers. This is consistent with Fredriksson, Ockert, & Oosterbeek (forthcoming) who find that the imputed wage effect of class-size based on test scores are substantially smaller than the direct estimates.

VIII *Conclusions*

This paper presents a two-factor model such that all student outcomes are a function of their cognitive and non-cognitive ability. The model shows that one can use a variety of short run outcomes to construct a measure of student cognitive and non-cognitive ability and use this to estimate a teacher's predicted effect on long run outcomes. Using longitudinal survey data, evidence indicates that non-cognitive ability (an index of non-test score socio-behavioral outcomes in 8th grade) is associated with sizable improvements in adult outcomes, and that non-cognitive ability is at least as important a determinant of adult outcomes as cognitive ability.

Using administrative data with students linked to individual teachers, 9th grade English and Algebra teachers have economically meaningful effects on test scores, absences, suspensions, on time 10th grade enrollment, and grades. The results indicate that teachers have similarly sized effects on cognitive ability as they do on non-cognitive ability. A variety of additional tests suggest that these effects can be interpreted causally. Teacher effects on test scores and teacher effects on non-cognitive ability are weakly correlated such that 38 percent of all teachers in the top 25 percent of test score value-added are also in the bottom 25 percent of teachers at improving non-cognitive skills. This means that a large share of teachers thought to be highly effective based on test score performance will be no better than the average teacher at improving college going or wages. Further calculations indicate that test score based measures may understate the importance of teachers by as much as 70 percent.

This study highlights that a failure to account for the effect of educational interventions on non-cognitive skills can lead to biased estimates of the effect of such interventions on important long-run outcomes. The two-sample factor analytic framework put forth in this paper can be used in other settings to estimate the effects of educational interventions on both cognitive and non-cognitive ability. Finally, the results illustrate the large potential gains that may be associated with making policy decisions about educational interventions based on estimated effects on both cognitive and non-cognitive outcomes rather than just test scores alone.

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. 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.
3. Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the BigFive in late childhood. *Personality and Individual Differences, 34*(4), 645-664.

4. 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.
5. Borghans, L., Weel, B. t., & Weinberg, B. A. (2008). Interpersonal Styles and Labor Market Outcomes. *Journal of Human Resources, 43*(4), 815-58.
6. Bowles, S., Gintis, H., & Osborne, M. (2001). The Determinants of Earnings: A Behavioral Approach. *Behavioral Approach, 39*(4), 1137-76.
7. Brookhart, S. M. (1993). Teachers' Grading Practices: Meaning and Values. *Journal of Educational Measurement, 30*(2), 123-142.
8. Carneiro, P., Crawford, C., & Goodman, A. (2007). The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes. *CEE Discussion Papers 0092*.
9. Cascio, E., & Staiger, D. (2012). Knowledge, Tests, and Fadeout in Educational Interventions. *NBER working Paper Number 18038*.
10. 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.
11. Chetty, R., Friedman, J., & Rockoff, J. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. *unpublished manuscript*.
12. Deming. (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*(3), 111-134.
13. Deming, D. (2011). Better Schools, Less Crime? *The Quarterly Journal of Economics, 126*(4), 2063-2115.
14. Downey, D., & Shana., P. (2004). When Race Matters: Teachers' Evaluations of Students' Classroom Behavior. *Sociology of Education, 77*, 267-82.
15. 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.
16. 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.
17. Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica, 77*(3), 953-973.
18. Fredriksson, P., Ockert, B., & Oosterbeek, H. (forthcoming). Long-Term Effects of Class Size. *Quartlerly Journal of Economics*.
19. 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.
20. 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.
21. Heckman, J. (1999). Policies to Foster Human Capital. *NBER Working Paper 7288*.
22. Heckman, J. J., & Rubinstein, Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review, 91*(2), 145-49.
23. 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.

24. Heckman, J., Pinto, R., & Savelyev, P. (forthcoming). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*.
25. 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.
26. Howley, A., Kusimo, P. S., & Parrott, L. (2000). Grading and the ethos of effort. *Learning Environments Research, 3*, 229-246.
27. Jackson, C. K. (2012). Teacher Quality at the High-School Level: The Importance of Accounting for Tracks. *NBER Working Paper 17722*.
28. Jackson, C. K. (forthcoming). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics*.
29. 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.
30. Jencks, C. (1979). *Jencks, Christopher.* New York: Basic Books.
31. 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.
32. 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.
33. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Pearson.
34. Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER working paper 14607*.
35. Koedel, C. (2008). An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Department of Economics, University of Missouri Working Paper 0808*.
36. Koedel, C. (2008). Teacher Quality and Dropout Outcomes in a Large, Urban School District. *Journal of Urban Economics, 64*(3), 560-572.
37. 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.
38. 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.
39. Lucas, S. R., & Berends, M. (2002). Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking. *Sociology of Education, 75*(4), 328-348.
40. Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73*(2), 417-458.
41. Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*.
42. Sadker, D. M., & Zittleman, K. (2006). *Teachers, Schools and Society: A Brief Introduction to Education.* McGraw-Hill.
43. Waddell, G. (2006). Labor-Market Consequences of Poor Attitude and Low Self-Esteem in Youth. *Economic Inquiry, 44*(1), 69-97.

Tables

Table 1: *Correlations among Short-run Outcomes in the NELS-88 Data*

	NELS-88 Data						
	Ln of # Days Absent	Suspended	Grade Point Average	In 9th grade on time	Math Score in 8th Grade	Reading Score in 8th Grade	Non-cognitive Factor
Ln of # Days Absent	1						-0.519
Suspended	0.148	1					-0.666
Grade Point Average	-0.171	-0.243	1				0.644
In 9th grade on time	-0.051	-0.127	0.127	1			0.453
Math Score in 8th Grade	-0.075	-0.128	0.383	0.023	1		0
Reading Score in 8th Grade	-0.049	-0.126	0.342	0.024	0.803	1	0

	NCERDC Data						
	Ln of # Days Absent	Suspended	Grade Point Average	In 10th grade on time	Algebra Score in 9th Grade	English Score in 9th Grade	Non-cognitive Factor
Ln of # Days Absent	1						-0.753
Suspended	0.252	1					-0.379
Grade Point Average	-0.232	-0.192	1				0.622
In 10th grade on time	-0.167	-0.16	0.482	1			0.4303
Algebra Score in 9th Grade	-0.098	-0.13	0.592	0.321	1		0
English Score in 9th Grade	-0.082	-0.13	0.539	0.323	0.616	1	0

Note: The non-cognitive factor was uncovered using factor analysis and is a linear combination of all the non-test score outcomes. I used regression to remove any lingering linear association between the factor and test scores. The result was then standardized. Note that the non-cognitive factor in the NCERDC uses the weights derived from the NELS-88 data. However, the non-cognitive factor using weights derived from the NCERDC has correlation 0.98 with one derived using weights from the NELS-88. Also note that the correlation between math and English test scores and the cognitive factor (not shown) in the NELs data is 0.94 and 0.93 respectively while that in the NCERDC data is 0.91 and 0.87, respectively.

Table 2: *Relationship between 8th Grade Outcomes and Adult Outcomes in the NELS-88 Survey.*

	1	2	3	4	5	6	7	8	9
	Arrested, or Close Friend Arrested			Attend College			Log of Income in 1999		
Ln of # Days Absent	**0.0146****	**0.0145****		**-0.0181****	**-0.0182****		-0.0101	-0.00965	
	[0.00528]	**[0.00528]**		**[0.00498]**	**[0.00496]**		[0.00957]	[0.00956]	
Grade Point Average	**-0.0389****	**-0.0451****		**0.133****	**0.117****		**0.103****	**0.0924****	
	[0.00574]	**[0.00617]**		**[0.00547]**	**[0.00592]**		**[0.0104]**	**[0.0111]**	
Repeat 8th Grade	0.00506	0.0014		**-0.168****	**-0.177****		0.0493	0.042	
	[0.0392]	[0.0393]		**[0.0406]**	**[0.0407]**		[0.0641]	[0.0640]	
Suspended	**0.130****	**0.131****		**-0.165****	**-0.162****		**-0.0649****	**-0.0648****	
	[0.0138]	**[0.0138]**		**[0.0135]**	**[0.0135]**		**[0.0222]**	**[0.0221]**	
Math Score in 8th Grade		0.0109	-0.00574		**0.0325****	**0.0767****		**0.0508****	**0.0555****
		[0.00709]	[0.00421]		**[0.00605]**	**[0.00406]**		**[0.0128]**	**[0.00779]**
Reading Score in 8th Grade		0.00347			0.00299			**-0.0315***	
		[0.00695]			[0.00602]			**[0.0124]**	
Non-cognitive Factor			**-0.0559****			**0.107****			**0.0597****
			[0.00440]			**[0.00423]**			**[0.00732]**
Demographics	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	10,792	10,792	10,792	10,792	10,792	10,792	9,956	9,956	9,956

Robust standard errors in brackets. + significant at 10%; * significant at 5%; ** significant at 1%

Notes: All models include controls for household income while in 8th grade, whether English is the students primary language at home, whether the student lives with their mother, whether the student lives with their father, the mothers highest level of education, the fathers highest level of education, family size, student race and student gender. Regressions of log income are median regression (unconditional qunatiles) rather than OLS to account for any bias due to very low levels of earnings. Also, note that all the independent variables were collected in 1988 when respondent were in 8^{th} grade while the outcomes were collected in 1999 when respondents were approximately 25 years old. Models that include both 8^{th} grade math and reading scores suffer from collinearity so only effects based on 8^{th} grade math scores are reported.

Table 3: *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 10^{th} grade	0.856	(0.351)	(0.305)	(0.339)
Observations			348547	

Notes: These summary statistics are based on student 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 so that the missing category is "missing 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 under Different Models*

		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
School Effects with Peer Characteristics	Algebra Score 9th	**0.12**	**0.000**	**0.128**	**0.112**	**0.068**	**0.000**	**0.079**	**0.054**
	English Score 9th	0.029	0.056	0.043	0.000	**0.049**	**0.000**	**0.059**	**0.037**
	Suspended	**0.023**	**0.001**	**0.029**	**0.015**	**0.032**	**0.000**	**0.036**	**0.028**
	Log of # Absences	**0.138**	**0.000**	**0.154**	**0.119**	**0.132**	**0.000**	**0.143**	**0.12**
	GPA	**0.095**	**0.000**	**0.103**	**0.086**	**0.081**	**0.000**	**0.089**	**0.072**
	On time enrollment	**0.058**	**0.000**	**0.064**	**0.052**	**0.047**	**0.000**	**0.053**	**0.041**
	Non-cognitive factor	**0.16**	**0.000**	**0.17**	**0.148**	**0.135**	**0.000**	**0.145**	**0.125**
Track-by-School and Year Effects	Algebra Score 9th	**0.1**	**0.000**	**0.107**	**0.093**	**0.039**	**0.005**	**0.051**	**0.019**
	English Score 9th	0.013	0.842	0.034	0.000	**0.043**	**0.000**	**0.05**	**0.035**
	Suspended	0.009	0.645	0.02	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	**0.108**	**0.000**	**0.121**	**0.094**	**0.097**	**0.000**	**0.109**	**0.084**
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	**0.03**	**0.003**	**0.04**	**0.014**
	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**	**0.000**
	GPA	**0.045**	**0.000**	**0.057**	**0.028**	**0.027**	**0.075**	**0.039**	**0.000**
	On time enrollment	**0.025**	**0.002**	**0.033**	**0.012**	**0.024**	**0.000**	**0.031**	**0.015**
	Non-cognitive factor	**0.049**	**0.049**	**0.07**	**0.003**	**0.049**	**0.006**	**0.079**	**0.025**

Notes: The estimated covariances are computed by taking the classroom level residuals from equation X and computing the covariance of mean residuals across classrooms for the same teacher. Specifically, I pair each classroom with a randomly chosen other 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 other classroom for the 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 which 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.

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	7	8
	Algebra Teachers (137,734 student observations)				English Teachers (284,560 student observations)			
	Predicted Algebra Scores	Algebra Scores	Predicted Non cognitive Factor	Non cognitive factor	Predicted English Scores	English Scores	Predicted Non cognitive Factor	Non cognitive factor
Estimated Teacher Effect (out of sample)	0.0214	**0.283****	-0.00583+	**0.0603***	0.0414+	**0.214****	0.00610+	**0.0468***
	[0.0210]	**[0.0329]**	[0.00349]	**[0.0263]**	[0.0237]	**[0.0230]**	[0.00331]	**[0.0216]**
Mean Estimated Teacher Effects (out of sample)	0.0663	**0.279****	-0.0061	**0.0628***	-0.0172	**0.230****	-0.0073	**0.0782****
	[0.0588]	**[0.0331]**	[0.0115]	**[0.0284]**	[0.0544]	**[0.0264]**	[0.00982]	**[0.0250]**

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

Notes: All models present the estimated coefficient on the own teachers estimated effect for the same outcome and the mean of the teacher effects at the school-track-year level. Both models are estimated simultaneously on the same sample using systems OLS regression. Models that cluster the standard errors at the teacher level or the track level yield smaller standard errors and are therefore less conservative than those reported.

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

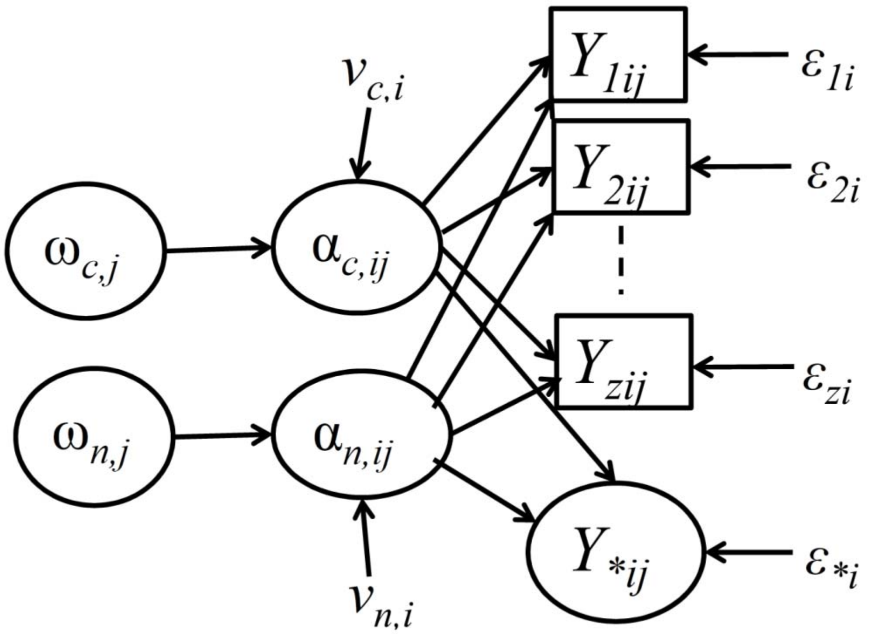
	Algebra Score	English Score	Suspended	Log of # Absences	GPA	On time enrollment in 10th grade	Non-cognitive factor
Algebra FX	1	0.0238	0.0077	0.025	**0.1068**	0.0503	0.0091
Non-cog	0.0091	0.0291	**0.1623**	**0.3747**	**0.5628**	**0.2597**	1

	Algebra Score	English Score	Suspended	Log of # Absences	GPA	On time enrollment in 10th grade	Non-cognitive factor
English FX	0.012	1	0.0089	0.0073	0.0491	0.0665	0.0186
Non-cog	0.038	0.0186	**0.2633**	**0.4016**	**0.5649**	**0.3283**	1

This table 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.

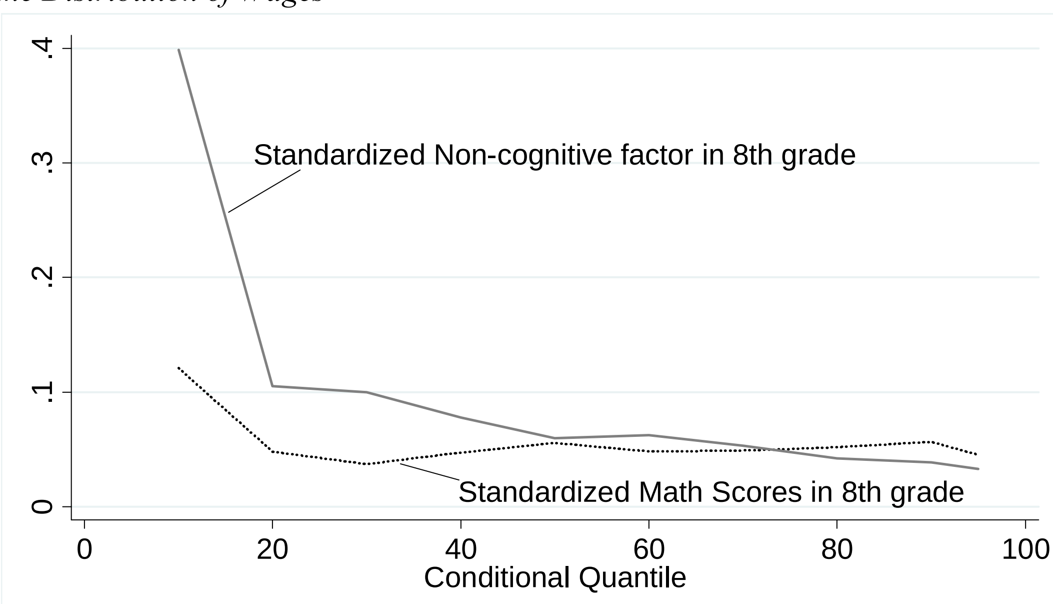
Figures

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



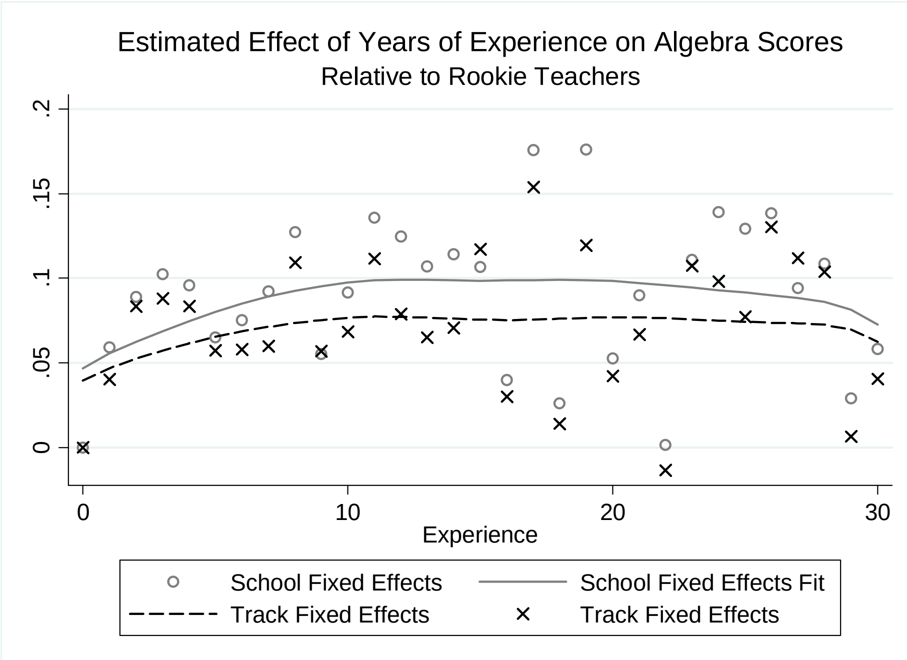
Note: An arrow from a to b indicates that variable b is a linear function of variable a. Square boxes denote observed variables; while ovals denote unobserved or latent variables.

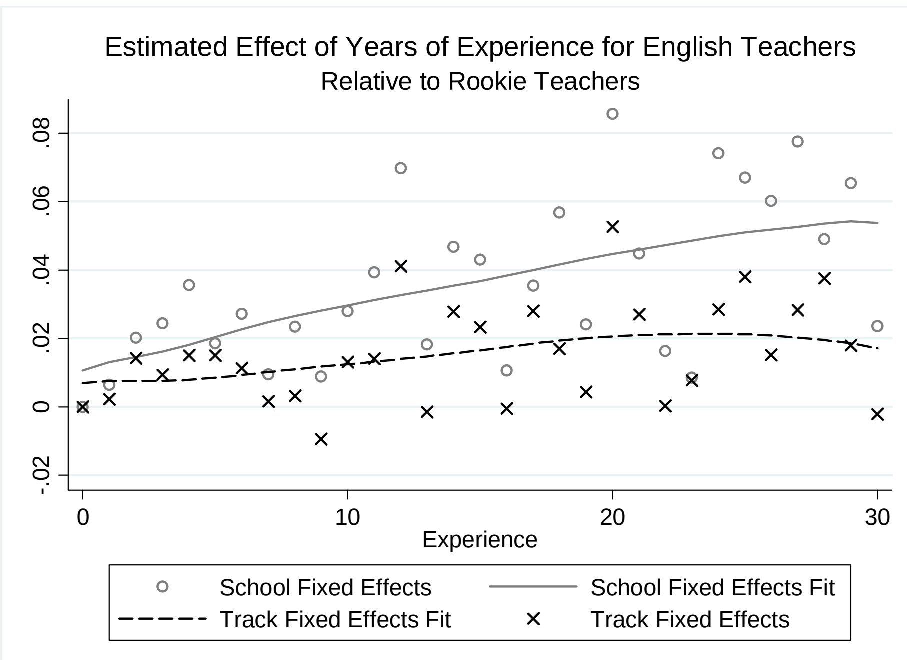
Figure 2: *The Marginal Effect of Cognitive and Non-cognitive Skills at Different Points of the Distribution of Wages*



Note: This figure presents estimates from models that control for demographics. The effects are even more pronounced when they are not included.

Figure 3: *The Marginal Effect of Teacher Experience under Different Models*





These figures present the estimated coefficients on indicator variables denoting each year of experience on English and Math test scores. The figures show the estimated point estimates for each year of experience and a lowess fit of the point estimates for models with school fixed effects, and track-by-school fixed effects.

Figure 4: *Relationship between Teacher Effects on Test Scores and Non-cognitive Factor*

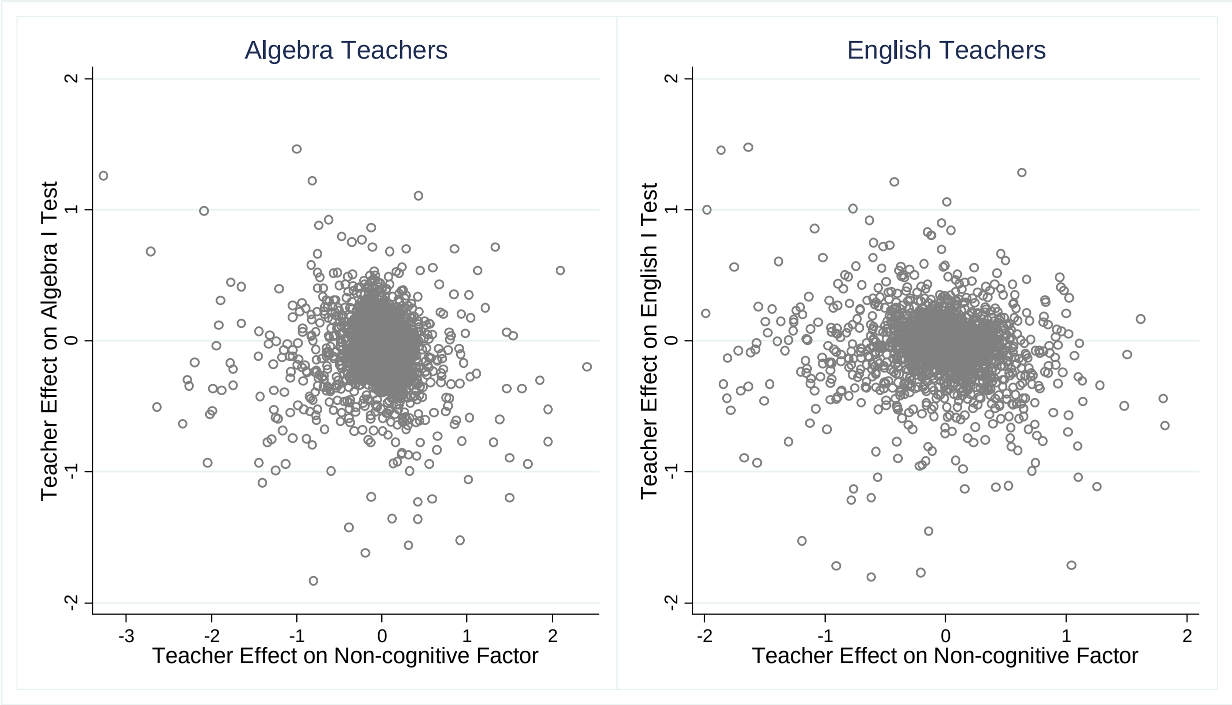
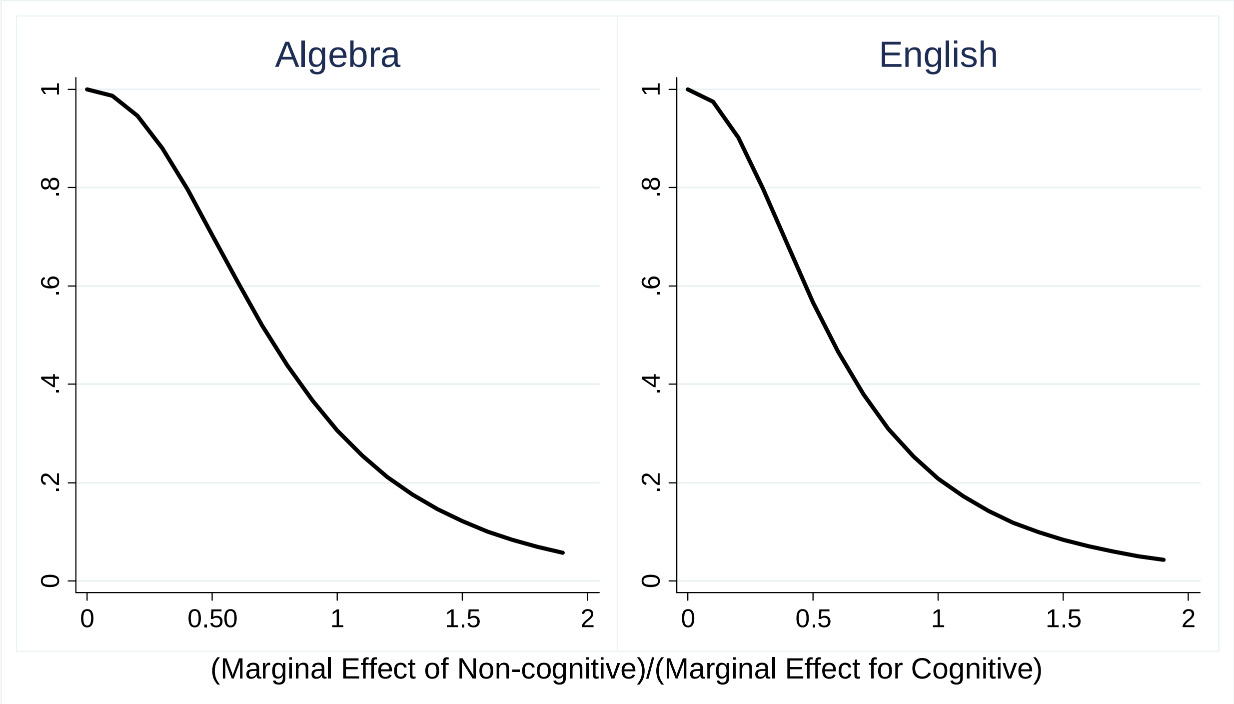


Figure 5: *The Proportion of the Teacher Effect on Predicted Adult Outcomes Explained by Test Score Value Added for Different Relative Weights on Non-cognitive Ability.*



Appendix

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 Track-Year-School Cell	English	Algebra
1	63.37	51.07
2	18.89	26.53
3	9.12	11
4	5.6	6.38
5	3.03	3.25
6	0	1.77

Note: This is after removing singleton tracks.

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 The 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 (2011) I group students into classrooms based in 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 the 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), 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 that error due to the fuzzy matching algorithm does not generate any of the empirical findings.