Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies

Matthew A. Kraft
Brown University

February 2017

Abstract

I exploit the random assignment of class rosters in the Measures of Effective Teaching Project to estimate teacher effects on students' performance on cognitively demanding open-ended tasks in math and reading, as well as their growth mindset, grit, and effort in class. I find large teacher effects across this expanded set of student outcomes, but weak relationships between these effects and multiple measures used in new teacher evaluation systems including effects on state standardized tests. These findings suggest that high-stakes evaluation decisions do not fully consider the degree to which teachers are developing students' complex cognitive skills or social-emotional competencies.

JEL No. H0, I2, J24

Correspondence regarding the paper can be sent to Matthew Kraft at mkraft@brown.edu. PO Box 1983, Brown University, Providence RI, 02912. This research was generously supported by the William T. Grant Foundation and the Brown University Undergraduate Training and Research Award program. I thank seminar participants at Brown, Harvard, Stanford, the University of Connecticut and the Federal Reserve Banks of Boston and New York for their helpful comments. I am grateful to Sarah Grace for providing exceptional and extensive research assistance on earlier versions of the paper as well as to Bruna Lee, Dylan Hogan, and Harry Neuert. All views and errors are my own.

1. Introduction

Although it is well established that teachers have large effects on student achievement, current evidence is largely limited to student performance on state standardized tests (Rockoff 2004; Hanushek and Rivkin 2010, Chetty, Friedman and Rockoff 2014a). These tests have typically measured students' core content knowledge as well as basic literacy and numeracy skills using multiple-choice questions. However, many of the ways in which teachers affect students' long term outcomes such as earnings (Chetty, Friedman and Rockoff 2014b) may be through their influence on skills and competencies not captured on state standardized tests (Bowles, Gintis and Osborne 2001). Using data from the Project STAR class-size experiment, Chetty and his colleagues (2011) estimated that only 1/5 of the total variation in class effects on earnings operated through effects on scores from multiple-choice tests. Jackson (2016) found that pairing teacher effects on state tests with effects on an index of behavioral outcomes more than doubles the predictive power of these effects for high school graduation as well as for early indicators of college-going.

This paper provides new evidence on the degree to which teachers affect a broad set of complex cognitive skills and social-emotional competencies using data across six large school districts collected by the Measures of Effective Teaching (MET) Project. Past MET Project reports have primarily focused on developing a composite measure of teacher effectiveness for forecasting effects on student achievement (Kane and Staiger, 2012) and validating this measure using random assignment (Kane et al. 2013).[1] Existing research linking teacher effects to outcomes other than traditional standardized assessments has examined three general outcome

[1] Included in these reports are estimates of teacher effects on open-ended cognitively demanding tests in a value-added framework (Kane and Cantrell, 2010; Tables 4 & 5) and estimates of the causal relationship between their composite measure of teacher effectiveness and students' social-emotional competencies (Kane et al. 2013; Table 14).

types: observable behavioral and schooling outcomes such as absences, suspensions, grades, grade retention and high-school graduation (Jackson 2016, Gershenson 2016, Koedel 2008, Ladd and Sorensen 2017), student self-reported attitudes and behaviors including motivation and self-efficacy in math, happiness and behavior in class, and time spent reading and doing homework outside of school (Blazar and Kraft 2017, Ladd and Sorensen 2017, Ruzek et al. 2014), and teacher assessments of students' social and behavioral skills (Chetty et al. 2011, Jennings and DiPrete 2010). These studies almost uniformly find teacher effects on non-test-score outcomes, often of comparable or even larger magnitude than effects on achievement (e.g. Blazar and Kraft 2017, Jennings & DiPrete 2010, Jackson, 2016).

The MET Project's scale and unique set of student measures combined with its experimental design allow me to make several important contributions to this literature. In addition to collecting administrative and achievement data, MET researchers administered two supplemental achievement tests with open-ended questions that were designed to be more direct measures of students' critical thinking skills and problem-solving skills on open-ended tasks. In the second year of the study, students also completed a questionnaire that included scales for measuring their grit (Duckworth and Quinn 2009) and growth mindset (Dweck 2006), two widely-publicized social-emotional competencies that have received considerable attention from policymakers and the media in recent years.[2] The survey also included a class-specific measure of effort which allows me to compare teacher effects on both global and domain-specific social-emotional measures. These data allow me to present the first estimates of students' ability to perform complex tasks in math and reading as well as on students' grit, growth mindset and

[2] Paul Tough's best-selling book *How Children Succeed* propelled grit into the national dialogue about what schools should be teaching. The White House has convened meetings on the importance of "Academic Mindsets" (Yeager et al., 2013) and the Department of Education has commissioned a paper on "Promoting Grit, Tenacity, and Perseverance" (Shechtman, 2013).

effort in class. It also allows me to present among the first direct evidence of the relationship between teacher effects on state tests, open-ended cognitively demanding assessments, and social-emotional competencies.

In the second year of the MET Project, a subset of teachers participated in an experiment where researchers randomly assigned student rosters among sets of volunteer teachers in the same grades and schools. This design provides the opportunity to identify teacher effects on outcomes other than standardized state tests without relying on a strong conditional independence assumption. To date, the literature has relied exclusively on a covariate adjustment approach with varying combinations of fixed effects to resolve the non-random sorting of students to teachers. Although growing evidence suggests this approach can produce unbiased, although imprecise, teacher effects on standardized achievement tests (Chetty et al. 2014a, Kane et al. 2013), it is not clear whether these results hold for other outcomes. Finally, I provide among the first evidence on whether measures used in high-stakes teacher evaluation systems including classroom observations, principal ratings, and student surveys reflect teacher effects on complex cognitive skills and social-emotional competencies.

The MET Project data also present several limitations that are relevant for these analyses. The single year of experimental data combined with my focus on general education elementary classrooms complicates my ability to isolate teacher effects from peer effects and transitory shocks (Chetty et al., 2011). Blazar & Kraft (2017) compared teacher effects on students' attitudes and behaviors with and without allowing for class-specific effects and found that estimates that do not remove class-specific peer effects and shocks are inflated by approximately 15%. I present estimates both with and without peer-level controls to provide an approximate bounds for teacher effects. Controlling for peer-level covariates in cross-sectional models such as

the one used in this paper over-attributes classroom effects to peers resulting in conservative estimates of the true magnitude of teacher effects (Kane et al. 2013; Thompson, Guarino, and Wooldridge 2015). Throughout the paper I refer to my estimates as teacher effects while recognizing that the data do not allow me to definitively separate the joint effect of teachers, peers, and shocks.

I am also unable to test the predictive validity of estimated teacher effects on complex cognitive skills and social-emotional competencies using longer-term outcomes following Jackson (2016). Such analyses using the MET data are not possible because the MET Project focused on teachers and, thus, did not collect panel data on students. I instead leverage the nationally representative Educational Longitudinal Survey to illustrate the predictive validity of self-report scales that are close proxies for measures of grit and growth mindset on a range of educational, economic, personal and civic outcomes and review the causal evidence on interventions targeting these competencies.

Leveraging the MET class-roster randomization design, I find teacher effects on standardized achievement in math and English Language Arts (ELA) that are similar in magnitude to prior analyses of the MET data (Kane & Cantrell 2010) and the broader value-added literature (Hanushek and Rivkin 2010). I also find teacher effects of comparable magnitude on students' ability to perform complex tasks in math and ELA, as measured by cognitively demanding open-ended tests. Teacher effects on students' social-emotional competencies differ in magnitude, with the largest effects on growth mindset, effort in class and the perseverance subscale of grit. Comparing the effects of individual teachers across these outcomes reveals that teachers who are most effective at raising student performance on standardized tests are not consistently the same teachers who develop students' complex

cognitive abilities and social-emotional competencies. While teachers who add the most value to students' performance on state tests in math do also appear to strengthen their analytic and problem-solving skills, teacher effects on state ELA tests are only moderately correlated with teacher effects on open-ended tests in reading. Successfully teaching more basic reading comprehension skills does not appear to translate consistently to the ability to interpret and respond to texts.

I find that teacher effects on social-emotional measures are only weakly to moderately correlated with effects on state achievement tests and more cognitively demanding open-ended tasks, even after adjusting for differential reliability in the measures. These findings suggest that teacher effectiveness differs across multiple dimensions. I then examine the relationship between estimated teachers effects with performance measures commonly incorporated into high-stakes teacher evaluation systems. I find little evidence that classroom observations, principal ratings, or student surveys are serving to capture teacher effects on this broader set of outcomes. I conclude by discussing the implications of my findings for research, policy and practice.

2. Schooling, Skills and Competencies

2.1 Complex Cognitive Skills

A growing number of national and international organizations have identified complex cognitive abilities as essential skills for the workplace in the modern economy (National Resource Council 2012; OECD 2013). Psychologists and learning scientists define complex cognitive skills as a set of highly interrelated constituent skills that support cognitively demanding processes (Van Merriënboer and Jeroen 1997). These skills allow individuals to classify new problems into cognitive schema and then to transfer content and procedural

knowledge from familiar schema to new challenges. Examples include writing computer programs, directing air traffic, engineering dynamic systems, or diagnosing sick patients.

Researchers and policy organizations have referred to these abilities using a variety of different terms including "21st Century Skills," "Deeper Learning," "Critical-Thinking" and "Higher-Order Thinking." State standardized achievement tests in mathematics and reading rarely include items designed to assess these abilities. A review of standardized tests used in 17 states judged as having the most rigorous state assessments found that 98% of items on math tests and 78% of items on reading tests only required students to recall information and demonstrate basic skills and concepts (Yuan and Le 2012). Open-ended ELA questions on state tests were substantially more likely to be judged as cognitively demanding assessments of "deeper learning." However, while open-ended test items in math required students to move beyond recall, they rarely required students to perform extended unstructured problems.

To date, empirical evidence linking teacher and school effects to the development of students' complex cognitive skills remains very limited. Researchers at RAND found that students who had more exposure to teaching practices characterized by group work, inquiry, extended investigations, and emphasis on problem-solving performed better on the open-ended math and science tests designed to asses students' decision making abilities, problem-solving skills, and conceptual understanding (Le et al. 2006). Using a matched-pair design, researchers at American Institutes for Research found that students attending schools that were part of a "deeper learning" network outperformed comparison schools by more than one tenth of a standard deviation in math and reading on the PISA-Based Test for Schools (PBTS) —a test that assesses core content knowledge and complex problem-solving skills (Zeiser et al 2014).

2.2 Social-Emotional Competencies

Social-emotional competencies (or social and emotional learning) is a broad umbrella term used to encompass an interrelated set of cognitive, affective and behavioral abilities that are not commonly captured by standardized tests. Although sometimes referred to as non-cognitive skills, personality traits, or character skills, these competencies explicitly require cognition, are not fixed traits, and are not intended to suggest a moral or religious valence. They are skills, attitudes and mindsets which can be developed and shaped over time (Duckworth and Yeager 2015). Regardless of the term used, mounting evidence documents the strong predictive power of competencies other than performance on cognitive tests for educational, employment, health and civic outcomes (Almlund et al. 2011; Borghans et al. 2008; Moffitt et al. 2011).

Two seminal experiments in education, the HighScope Perry Preschool Program and Tennessee Project STAR, documented the puzzling phenomenon of how the large effects of high-quality early-childhood and kindergarten classrooms on students' academic achievement faded out over time, but then reappeared when examining adult outcomes such as employment and earnings as well as criminal behavior. Recent re-analyses of these experiments suggest that the long-term benefits of high-quality pre-K and kindergarten education were likely mediated through increases in students' social-emotional competencies (Heckman, Pinto and Savelyev 2013; Chetty et al. 2011).

3. Research Design

3.1 The MET Project

The MET Project was designed to evaluate the reliability and validity of a wide range of performance measures used to assess teachers' effectiveness. The study tracked approximately 3,000 teachers from across six large public school districts over the 2009-10 and 2010-11 school

years.[3] These districts included the Charlotte-Mecklenburg Schools, the Dallas Independent Schools, the Denver Public Schools, the Hillsborough County Public Schools, the Memphis Public Schools, and the New York City Schools. Across districts there is substantial variation in the racial composition of students where African-American, Hispanic and white students each comprise the largest racial/ethnic group in at least one district.

In the second year of the study, MET researchers recruited schools and teachers to participate in a classroom roster randomized experiment. Of those 4th and 5th grade general education teachers who participated in the first year and remained in the study in the second year, 85% volunteered for the randomization study and were eligible to participate. Participating principals were asked to create classroom rosters that were "as alike as possible in terms of student composition" in the summer of 2010 (Bill and Melinda Gates Foundation 2013, p. 22). They then provided these rosters to MET researchers to randomize among volunteer teachers in the same schools, subjects and grade levels.[4] The purpose of this randomization was to eliminate potential bias in teacher effect estimates caused by any systematic sorting of teachers and students to specific classes within schools.

I focus my empirical analyses on the effect of general education elementary classrooms to minimize the potential confounding when students are taught by multiple teachers and outcomes are not class-specific. Almost 8,000 elementary school students (n=7,999) were included on class rosters created for general elementary school teachers by principals. Similar to Kane et al. (2013), I find substantial attrition among the 4th and 5th grade students who were included in the roster randomization process; 38.6% of students on these rosters were not taught

[3] Detailed descriptions of the MET data are available at www.metproject.org.
[4] Detailed descriptions of the randomization design and process can be found in Kane et al. (2013) and the Measures of Effective Teaching User Guide (Bill & Melinda Gates Foundation, 2013).

by any teachers who participated in the MET Project data collection in 2010-2011 and thus are censored from the MET dataset. Much of this attrition is due to the randomization design, which required principals to form class rosters before schools could know which students and teachers would remain at the school. Following random assignment, some students left the district, transferred to non-participating schools, or were taught by teachers who did not participate in the MET study. Some participating teachers left the profession, transferred schools or ended up teaching different classes within their schools than originally anticipated. I present several analyses examining randomization balance in the analytic sample in section 4.1 and find that this attrition does not compromise the internal validity of the analyses to a great degree.

I construct the analytic sample to include only students in 4th and 5th grades who 1) were included in the roster randomization process 2) were taught by general education teachers who participated in the randomization study, and 3) have valid lagged achievement data on state standardized tests in both math and ELA. These restrictions result in an analytic sample of 4,151 students and 236 general education teachers. Further restricting the analytic sample to be a balanced panel where students have valid data for all outcomes would reduce the sample to 2,907 students. In analyses available upon request, I confirm that the primary results are unchanged when using this smaller balanced sample. I present descriptive statistics on the students and teachers in the analytic sample in Table 1. The sample closely resembles the national population of students attending public schools in cities across the United States but with a slightly larger percentage of African-American students and smaller percentage of white and Hispanic students: 36% are African-American, 29% are Hispanic, 24% are white, and 8% are Asian. Over 60% of students qualify for free or reduced-price lunch (FRPL) across the sample. The 4th and 5th grade general education elementary school teachers who participated in the MET Project

randomization design are overwhelmingly female and substantially more likely to be African American compared to the national labor market of public school teachers. Teacher experience varies widely across the sample, and half of all teachers hold a graduate degree.

3.2 Standardized State Tests

The MET dataset includes end-of-year achievement scores on state standardized tests in math and ELA, as well as scores from the previous year. Multiple-choice items were the primary question format used on the 4th and 5th grade state math and ELA tests administered in the six districts in 2011. State testing technical manuals suggest that the vast majority of items on these exams assessed students' content knowledge, fundamental reading comprehension and basic problem-solving skills.[5] Reported reliabilities for these 4th and 5th grade tests in 2011 ranged between 0.85-0.95. In order to make districts' scaled scores comparable across districts, the MET Project converted these scores into rank-based Z-scores.

3.3 Achievement Tests Consisting of Open-Ended Tasks

MET researchers administered two supplemental achievement tests to examine the extent to which teachers promote high-level reasoning and problem solving skills. The cognitively demanding tests, the Balanced Assessment in Mathematics (BAM) and the Stanford Achievement Test 9 Open-ended Reading Assessment (SAT9-OE), consist exclusively of constructed-response items. The BAM was developed by researchers at the Harvard Graduate School of Education and consists of four to five tasks that require students to complete a series of open-ended questions about a complex mathematical problem and justify their thinking. Similar

[5] Out of the six state ELA exams, four consisted of purely multiple-choice items (FL, NC, TN, and TX), while two also included open response questions (CO and NY). Among the math exams, two were comprised of multiple choice questions only (TN and TX), three contain gridded response items that require students to complete a computation and input their answer (CO, FL, and NC), and one included several short and extended response questions (NY).

to the BAM, the SAT9-OE developed by Pearson Education consisted of nine open-ended questions about one extended reading passage that tested students' abilities to reason about the text, draw inferences, explain their thinking and justify their answers. I estimate internal consistency reliabilities of students' scores across individual items on the BAM and SAT9-OE of 0.72 and 0.85, respectively.

Little direct evidence exists about the predictive validity of the BAM and SAT9-OE assessments, in part, because these tests were never commercialized at scale. These assessments were chosen by MET Project researchers based on the primary criterion that they "provide[d] good measures of the extent to which teachers promote high-level reasoning and problem solving skills" (MET Project, 2009). Although format alone does not determine the cognitive demand of test items, a review of six major national and international assessments using Webb's Depth-of-Knowledge framework found that 100% of writing, 52% of reading and 24% math open-response items assessed strategic or extended thinking compared to only 32% of reading and 0% of math multiple-choice items (Yuan & Lee, 2014). Demand and wages for jobs that require these complex cognitive skills to perform non-routine tasks, often in combination with a range of interpersonal skills, have grown steadily in recent decades (Autor, Levy and Murnane 2003; Deming 2015; Weinberger 2014).

3.4 Social-Emotional Measures

Students completed short self-report questionnaires to measure their grit and growth mindset in the second year of the study. The scale used to measure grit was developed by Angela Duckworth to capture students' tendency to sustain interest in, and effort toward, long-term goals. Students responded to a collection of eight items (e.g., "I finish whatever I begin") using a five-category Likert Scale, where 1 = *not like me at all* and 5 = *very much like me*. I estimate

student scores separately for the two subscales that comprise the overall grit measure as presented in the original validation study (Duckworth & Quinn, 2009): 1) consistency of interest and 2) perseverance of effort (hereafter consistency and perseverance). This approach provides an important opportunity to contrast a global measure of perseverance with a class-specific measure of effort described below and distinguishes between conceptually distinct constructs that have an unadjusted correlation of 0.23 and a disattentuated correlation of 0.34 in the analytic sample.

The growth mindset scale developed by Carol Dweck measures the degree to which students' views about intelligence align with an incremental theory that intelligence is malleable, as opposed to an entity theory, which frames intelligence as a fixed attribute (Dweck, 2006). Students were asked to rate their agreement with three statements (e.g., "You have a certain amount of intelligence, and you really can't do much to change it") on a six-category Likert scale, where 1 = *strongly disagree* and 6 = *strongly agree*. I complement these global social-emotional measures with a class-specific measure of effort, constructed from responses to survey items developed by the Tripod Project for School Improvement. The scale consists of six items on which students are asked to respond to a descriptive statement about themselves using a 5-category Likert scale, where 1 = *totally untrue* and 5 = *totally true* (e.g. "In this class I stop trying when the work gets hard").

Reliability estimates of the internal consistency for growth mindset, consistency, perseverance and effort in class are 0.78, 0.66, 0.69, and 0.56 respectively. I construct scores on each of the measures following Duckworth and Quinn (2009) and Blackwell et al. (2007) by assigning point values to the Likert-scale responses and averaging across the items in each scale. I then standardize all three social-emotional measures in the full MET Project sample within

grade-level in order to account for differences in response scales and remove any trends due to students' age that might otherwise be confounded with teacher effects across grade levels. See Appendix A for the complete list of items included in each scale.

While a large body of evidence documents the predictive validity of social-emotional measures such as the Big Five, locus of control, and self-esteem (Almlund et al. 2011; Borghans et al. 2008; Moffitt et al. 2011), evidence for grit and growth mindset is more limited. Grit has been shown to be predictive of GPAs at an Ivy League school, retention at West Point, and performance in the Scripps National Spelling Bee, conditional on IQ (Duckworth et al. 2007; Duckworth and Quinn 2009). Grittier soldiers were more likely to complete an Army Special Operations Forces selection course, grittier sales employees were more likely to keep their jobs, and grittier students were more likely to graduate from high school, conditional on a range of covariates (Eskreis-Winkler et al. 2014). Middle school students who report having a high growth mindset have been found to have higher rates of math test score growth than students who view intelligence as fixed (Blackwell, Trzesniewski and Dweck 2007).

Given the lack of medium or long term outcomes in the MET data, I examine the predictive validity of social-emotional measures, conditional on standardized test scores, on students' educational attainment, labor market, personal and civic outcomes ten years later using the Educational Longitudinal Study (ELS). As predictors I use proxy measures of grit and growth mindset constructed from 10^{th} grade students' self-reported answers to survey items that map closely onto the perseverance of effort subscale of grit and provide a domain-specific measure of students' growth mindset in math. I create a composite measure of students' academic ability in math and reading based on students' scores on a multiple-choice achievement test administered by National Center for Education Statics (See Appendix B for details).

In Table 2 I report results from a simple set of OLS regression models where standardized measures of academic achievement, grit (perseverance) and growth mindset are included simultaneously with controls for students' race and gender, and level of parental education and household income. Although grit and growth mindset are generally weaker predictors of outcomes in adulthood compared to measures of academic achievement, the conditional relationships I estimate are still of meaningful economic magnitudes. For example, a one standard deviation increase in grit and growth mindset (0.61 and 0.73 scale points on a 4 point scale, respectively) is associated with $1,632 and $848 increases in annual employment income, respectively, as well as 5.8 and 1.1 percentage point increases in the probability a student has earned a bachelor's degree by age 26. Both grit and growth mindset are negatively associated with teen pregnancy and positively associated with civic participation. These conditional associations are likely conservative estimates of the predictive power of grit and growth mindset as they are not disattentuated for the lower reliability of survey-based measures and the measure of growth mindset is math-specific rather than the global measure used in the MET Project.

These analyses demonstrate that grit and growth mindset measures contain information that predicts outcomes in adulthood independent from academic ability, but they do not prove an underlying causal relationship. A growing number randomized control trials evaluating the effect of growth mindset interventions have documented causal effects on short to medium-term academic and behavioral outcomes (Yeager et al. 2014; Miu & Yeager 2015; Paunesku et al. 2015; Yeager et al. 2016). These studies demonstrate that growth mindsets interventions increased math and science GPA over several months (Yeager et al. 2014), satisfactory performance in high-school courses (Paunesku et al. 2015), and classroom motivation

(Blackwell, Trzesniewski and Dweck 2007) as well as decreased self-reported depressive symptoms (Miu and Yeager 2015) and aggressive desires and hostile intent attributions (Yeager et al. 2013).

The causal evidence on the effect of grit is more limited. Several small-scale field experiments document the short-term positive academic effects of mental contrasting strategies where students learn how to plan for and overcome obstacles for achieving their goal (Duckworth et al. 2011; Duckworth et al. 2013). A recent study found that teaching 4th grade students in Turkey about the plasticity of the human brain, the importance of effort, learning from failures, and goal setting improvement performance and persistence on objective tasks and grades (Sule et al. 2016). Together, these studies suggest that growth mindsets and grit are both malleable and likely causal determinants of important intermediary student outcomes for success in later life.

3.5 Achievement Tests, Performance on Open-Ended Tasks, and Social-Emotional Competencies

In Table 3, I present Pearson correlations across the eight outcomes measures. The clustered patterns of covariance evident in this table illustrate how each of these measures are not independent. Instead, these outcomes likely capture a more limited set of latent constructs. I provide the unadjusted correlations as well as technical details about how I disattenuate correlations for measurement error in Appendix C. The strongest relationships are between students' performance on state standardized tests across subjects (0.81) and students' math performance on the state tests and the open-ended test (0.81). This suggests that students who perform well on more-basic multiple-choice math questions tend to also perform well on more demanding open-ended math tasks. Student performance on state ELA tests and the SAT9-OE are correlated at 0.56, suggesting that state ELA tests are imperfect proxies for students' more

complex reasoning and writing skills. Correlations between social-emotional measures and state tests as well as open-ended tests are positive but of more moderate magnitude, ranging between 0.22 and 0.41. The pattern of correlations among the social-emotional measures themselves suggest that these scales may capture two distinct competencies: self-regulation and academic mindsets. Grit subscales (especially the perseverance subscale) and effort in class are moderately to strongly correlated and can both be characterized as measures of students' ability to self-regulate their behavior and attention.

3.6 Estimating the Variance of Teacher Effects

I begin by specifying an education production function to estimate teacher effects on student outcomes. A large body of literature has examined the consequences of different model specifications (Todd and Wolpin 2003; Kane and Staiger 2008; Koedel and Betts 2011; Guarino, Reckase, and Wooldridge 2015; Chetty, Friedman, and Rockoff 2014a). Typically, researchers exploit panel data with repeated measures of student achievement to mitigate against student sorting by controlling for prior achievement. The core assumption of this approach is that a prior measure of achievement is a sufficient summary statistic for all the individual, family, neighborhood, and school inputs into a student's achievement up to that time. Models also commonly include a vector of student characteristics, averages of these characteristics and prior achievement at the classroom level, and school fixed effects (see Hanushek and Rivkin 2010).

Researchers often obtain the magnitude of teacher effects from these models by quantifying the variance of teacher fixed effects, $\hat{\sigma}^2_{\tau_{FE}}$, or "shrunken" Empirical Bayes (EB) estimates, $\hat{\sigma}^2_{\tau_{EB}}$. EB estimates are a weighted sum of teachers' estimated effect, $\hat{\tau}_j$, and the average teacher effect, $\bar{\tau}$.

$$(1) \qquad E\left[\tau_j \mid \hat{\tau}_j\right] = \left(1-\lambda_j\right)\bar{\tau} + \left(\lambda_j\right)\hat{\tau}_j \qquad \text{where} \qquad \lambda_j = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_{\bar{\varepsilon}_j}}$$

Here the weights are determined by the reliability of each estimate, where λ_j is the ratio of true teacher variation to total teacher variance. However, fixed effects estimates are biased upward because they conflate true variation with estimation error. EB estimates are biased downward proportional to the size of the measurement error in the unshrunken estimates (see Jacob and Lefgren 2005, Appendix C). The true variance of teacher effects, σ_τ^2, is bounded between the fixed effect and EB estimators (Raudenbush and Bryk 2002).

(2) $$\hat{\sigma}^2_{\tau_{FE}} > \sigma^2_\tau > \hat{\sigma}^2_{\tau_{EB}}$$

Following Nye et al. (2004) and Chetty et al. (2011), I estimate the magnitude of the variance of teacher effects using a direct, model-based estimate derived via restricted maximum likelihood estimation. I assume a Gaussian data generating process which appears well justified in the data for state and open-ended tests and an appropriate approximation for social-emotional measures. This approach is robust to the differences in reliabilities across student outcomes — assuming classical measurement error — because it simultaneously models systematic unexplained variance across teachers as well as idiosyncratic student-level variance. It produces both a maximally efficient and consistent estimator for the true variance of teacher effects.

To arrive at this model-based estimate, I specify a multi-level covariate-adjustment model as follows:

(3) $$Y_{ij} = \alpha_{dg}(f(A_{i,t-1})) + \delta X_i + \beta \bar{A}_{j,t-1} + \theta \bar{X}_j + \pi_{sg} + \varepsilon_{ij}$$

$$\text{where} \quad \varepsilon_{ij} = \tau_j + \epsilon_i$$

Here, Y_{ij}, is a given outcome of interest for student *i* in district *d*, in grade *g*, with teacher *j* in school *s* in year *t*. Across all model specifications, I include a cubic function of students' prior year achievement on state standardized tests ($A_{i,t-1}$), in both mathematics and ELA which I allow to vary across districts and grades by interacting the linear lagged test score terms with district-by-grade fixed effects. I also include a vector of controls for observable student characteristics (X_i). Student characteristics include indicators that control for a student's gender, age, race, FRPL, English proficiency status, special education status, and participation in a gifted and talented program.[6]

I supplement these administrative data with additional student-level controls constructed from survey data collected by the MET Project. These include controls for students' self-reported prior grades, the number of books in their homes, the degree to which English is spoken at home, and the number of computers in their homes.[7] Both theory and prior empirical evidence have shown that grades reflect students' cognitive skills as well as social-emotional competencies such as grit and effort (Bowen, Chingos, McPherson 2009). I find that this measure of grades is positively correlated with social-emotional measures even when controlling for prior achievement in math and ELA. Partial correlations in the analytic sample range from 0.04 with growth mindset to 0.22 with perseverance. I include randomization block fixed effects (π_{sg}) to account for the block randomized design and restrict the estimation samples to exclude any classrooms where less than five students had valid outcome measures.

[6] Data on FRPL was not provided by one district. We account for this by including a set of district-specific indicators for FRPL and imputing all missing data as zero.
[7] We impute values of zero for students with missing survey data and include an indicator for missingness.

In additional models, I attempt to remove peer effects by controlling for a rich set of average classroom covariates.[8] These covariates include the average prior achievement in a student's class in both subjects ($\bar{A}_{j,t-1}$) as well as average student characteristics (using both administrative and survey data) in a students' class ($\bar{X}_j$). I present models both with and without peer effects to provide an informal upper and lower bounds on the true magnitude of teacher effects. Estimates of the magnitude of teacher effects from our models are likely to be biased upwards when peer-level controls are omitted and biased downward when they are included if peer effects are correlated with teacher quality (Kane et al. 2013; Thompson, Guarino, and Wooldridge 2015).

I allow for a two-level error structure for ε_{ij} , where τ_j represents a teacher-level random effect and ϵ_i is an idiosyncratic student-level error term. I obtain an estimate of the true variance parameter, $\hat{\sigma}_\tau^2$, directly from the model through maximum likelihood estimation. I specify τ_j in two different ways – as students' actual teachers and their randomly assigned teachers. Modeling the effects of students' actual teachers may lead to potentially biased estimates due to noncompliance with random assignment. Among those students in the analytic sample, 28.1% are observed with non-randomly assigned teachers. For this reason I include a rich set of administrative and survey-based controls. I further address the potential threat of non-compliance by exchanging the precision of actual-teacher estimates for the increased robustness of specifying τ_j as students' randomly assigned teachers. Estimates from this approach are analogous to Intent-to-Treat effects (ITT).

[8] We calculate peer characteristics based on all students who were observed in a teacher's classroom, regardless of whether they were included in the classroom roster randomization process or not.

4. Findings

4.2 Post-Attrition Balance Tests

I conduct two tests to assess the degree to which student attrition from the original randomized classroom rosters poses a threat to the randomization design. I begin by testing for balance in students' average characteristics and prior achievement across classrooms in the analytic sample. I do this by fitting a series of models where I regress a given student characteristic or measure of prior achievement, de-meaned within randomization blocks, on a set of indicators for students' randomly assigned teachers. In Table 4, I report F-statistics of the significance of the full set of randomly assigned teacher fixed effects. I find that, post-attrition, students' characteristics and prior achievement remain largely balanced within randomization blocks. For ten of these twelve measures, I cannot reject the null hypothesis that there are no differences in average student characteristics across randomly assigned teachers. However, I do find evidence of imbalance for students who participated in a gifted program or were an English language learner (ELL). This differential attrition likely occurred because gifted and ELL students were placed into separate classes with performance requirements or teachers who had specialized certifications.

I next examine whether there appears to be any systematic relationship between students' characteristics in the analytic sample and the effectiveness of the teachers to whom they were randomly assigned. In Table 5, I present results from a series of regression models in which I regress prior-year value-added scores of students' randomly assigned teachers on individual student characteristics and prior achievement. I do this for value-added estimates derived from both math and ELA state tests as well as the BAM and SAT9-OE exams in the prior academic

year.[9] Among the 48 different relationships I test, I find that only one is statistically significant at the 5% level. Post-attrition, students from low-income families are paired with randomly assigned teaches that have Math value-added scores that are, on average, 0.017 standard deviations (sd) higher on the state math exam in the prior year. This relationship is in the opposite direction from the type of sorting researchers are typically worried about, where more advantaged students are sorted to higher performing teachers. Even with the limited power for these tests, the magnitudes of these estimates, which are consistently less than 0.015 sd and never larger than 0.035 sd, are small relative to a standard deviation in the distribution of teacher effects in the non-experimental 2010 MET data (Math .226 sd; ELA .170 sd; BAM .211 sd; SAT9-OE .255 sd).

Together, these tests of post-attrition randomization balance across teachers suggest that the classroom roster randomization process did largely eliminate the systematic sorting of students to teachers commonly present in observational data (Kalogrides and Loeb 2013; Rothstein 2010). Although I observe some differential attrition across classrooms based on students' gifted and ELL status, there is little evidence that this attrition is related to teachers' effectiveness. To further examine this threat, I replicate my primary analyses in samples that exclude gifted and ELL students and find that the results are consistent with those reported below with the exception of both an absolute and relative increase in the magnitude of teacher effects on effort in class. Results are available upon request.

4.3 Teacher Effects – Maximum Likelihood Estimates

In Table 6, I present estimates of the standard deviation of teacher effects from a range of

[9] We use value-added estimates calculated by the MET researchers because the district-wide data necessary to replicate these estimates are not publically available. For more information about the value-added model specification see Bill & Melinda Gates Foundation, 2013.

models. Column 1 corresponds to the predominant school fixed effect specification in the teacher effects literature reviewed by Hanushek and Rivkin (2010). Consistent with prior studies, maximum likelihood estimates of the magnitude of teacher effects on state test scores are 0.16 sd in math and 0.14 sd in ELA. Using this baseline model, I also find teacher effects on the BAM and SAT9-OE tests of 0.14 sd and 0.16 sd, respectively. Finally, I find suggestive evidence of teacher effects on social-emotional measures ranging from 0.09 sd for consistency of interest (not statistically significant) to 0.20 for growth mindset.

In my preferred models with randomization-block fixed effects, I find strong evidence of teacher effects on students' complex task performance and social-emotional competencies, although the magnitude of these effects differ across measures. Columns 2 and 3 report results from models where I estimate teacher effects using students' actual teachers. In Columns 4 and 5, I exchange students' actual teachers with their randomly assigned teachers. For both specifications, I present results with and without peer effects. Comparing results across Columns 2 vs. 3 and 4 vs. 5 illustrates how the inclusion of peer-level controls somewhat attenuates my estimates by absorbing peer effects that were otherwise attributed to teachers. Focusing on estimates with students' actual teachers that condition on peer controls (Column 3), I find relatively similar estimates of the magnitude of teacher effects on most outcomes as in the baseline model. Teacher effects on growth mindset are attenuated (0.14 sd) and become similar in magnitude to effects on state tests.

As is common in field experiments in schools, there were some students who did not comply with the experimental design. In order to account for this non-compliance I estimate ITT effects of students' randomly assigned teachers. Results from these models are slightly attenuated given this non-compliance but remain consistent with estimates reported above.

Teacher effects on academic outcomes range from 0.12 sd on the BAM to 0.16 sd for the SAT9-OE. Teacher effects on consistency of interest do not achieve statistical significance, while effects on students' growth mindset (0.16 sd), perseverance (0.14) and effort in class (0.14) are of similar and even slightly larger magnitude than effects on achievement. Together, these results present strong evidence of meaningful teacher effects on students' social-emotional competencies and ability to perform complex tasks.

4.4 Comparing Teacher Effects across Outcomes

I investigate the nature of teacher skills by examining the relationships between individual teacher's effects across the eight outcomes of interest. In Table 7, I present Pearson correlations of Best Linear Unbiased Predictor estimates of teacher effects from the maximum likelihood (ML) model that uses students' actual teachers and includes peer controls (Column 3 of Table 6). Correlations among teacher effects from models using randomly assigned teachers produce a consistent pattern of results but are somewhat attenuated due to non-compliance. We present these results in Appendix D Table AD1.

Consistent with past research, I find that the correlation between general education elementary teachers' value-added on state math and ELA tests is large at 0.60 (Corcoran, Jennings and Beveridge 2012). Elementary teacher effects on state math tests also appear to be strongly related to their effects on the BAM (0.66). This suggests that teachers who are effective at teaching more basic computation and numeracy skills also appear to be developing their students' complex problem-solving skills in math. In contrast, teacher effects on state ELA exams are a poor proxy for teacher effects on more cognitively demanding open-ended ELA tests (0.25). In fact, teachers' value-added to student achievement on the SAT9-OE, which captures

students' ability to reason about and respond to an extended passage, is most strongly related to their effects on the similarly demanding open-ended math test (0.43).

I find that teacher effects on social-emotional measures are only weakly correlated with effects on both state standardized exams and exams testing students' performance on complex tasks. Among the four social-emotional measures, growth mindset has the strongest and most consistent relationship with teacher effects on state tests and complex task performance, with correlations ranging between 0.12 and 0.22. Teachers' ability to motivate their students' perseverance and effort is consistently a stronger predictor of teacher effects on students' complex task performance than on standardized tests scores. Finally, teacher effects across different social-emotional measures are far less correlated than teacher effects on student achievement across subjects. Effects on growth mindset are positively correlated with effects on students' consistency of interest (0.22), but unrelated to a teacher's ability to motivate students' perseverance and effort. Teacher effects on perseverance and effort in class are the only two social-emotional measures that appear to be capturing the same underlying ability among teachers, with a correlation of 0.61. This is important because it suggests that teacher effects on students' willingness to devote effort to their classwork may extend to other contexts as well.

I illustrate the substantial degree of variation in individual teacher effects across measures by providing a scatterplot of teacher effects on state math tests and growth mindset in Figure 1. This relationship captures the strongest correlation I observe between teacher effects on social-emotional competencies and state tests (0.22). A total of 43% of teachers in the sample have above average effects on one outcome but below average effects on the other (24% in quadrant II and 19% in quadrant IV). Only 31% of teachers have effects that are above average for both state math tests and growth mindset. The proportion of teachers who have above average effects on

both state math tests and other social-emotional measures is even lower.

4.5 Assessing Potential Bias in Teacher Effect Correlations

The pairwise correlations presented above are imperfect estimates of the true relationships between teacher effects on different outcomes, although the direction of potential bias is not obvious. Because these estimates are derived from the same sample of students for each general education teacher, unobserved student traits correlated with multiple outcomes will likely induce an upward bias. At the same time, noise in teacher effect estimates due to sampling error (the small number of students per teacher) and measurement error (imperfect reliability of student outcome measures) will likely bias estimates downward. I explore the potential magnitude of these biases below.

Past researchers have resolved the challenge of correlated errors by estimating these relationships using teacher effects from different classes or years. I am unable to estimate teacher effects across classes given the focus on general elementary school teachers. The single administration of the survey questions capturing students' self-reported grit and growth mindset in the second year of the study also prevents me from comparing teacher effects across years for these outcomes. I attempt to better understand this issue by comparing correlations both within and between class sections for teachers who taught multiple classes in the first year of the MET Project. I use teacher effects provided in the MET Project data that are derived from a standard value-added model (Kane & Cantrell, 2010). Estimates reported in Table 8 are available for five different outcomes but can only be compared within-subject given the available data.

I find that teacher effect correlations estimated from the same section of students are inflated relative to between-section teacher effects which eliminate the bias of common student shocks. The largest degree of inflation occurs for estimates between outcomes that are more

highly correlated such as state tests and the supplemental open-ended assessments administered by the MET project. Smaller correlations between teacher effects on achievement measures and students' self-reported effort in class are only slightly inflated if at all.

I examine the degree to which sampling error may attenuate estimated correlation coefficients by reestimating the correlation matrix within a common subsample of teachers that have a minimum of 15 students in their class (between 96 and 104 teachers across outcomes). I then repeatedly drop one student per teacher and reestimate the correlation matrix until the minimum class size reaches five students and plot the results in Figure 2. This figure illustrates how increasing the minimum sample size for estimating teacher effects slightly increases the magnitude for some, but not all, pair-wise correlations. The average increase in estimated correlations across all 28 pairwise correlations is only 0.003 suggesting that sampling error is unlike to substantially attenuate these estimates.

Finally, I can disattenuate these estimated correlations using an approach analogous to the Spearman (1904) adjustment described in Appendix C. I provide technical details for this procedure in Appendix E and report the estimated disattenuated correlations in Table AE2. The low estimated reliabilities of teacher effect estimates ranging between 0.51 and 0.56 result in almost a doubling of the magnitude of the unadjusted correlations. Give these low reliabilities, this adjustment should be viewed as extreme and providing only an upper bound estimate of the true correlations. For example, I find that correlations of approximately 0.60 and above are adjusted to be greater than 1, outside the possible range of correlation coefficients.

While it is difficult to know how these different biases interact, I interpret these findings to suggest that the low reliability of teacher effect estimates which attenuates correlations is not fully offset by the upward bias due to correlated errors from common student samples. I expect

the correlations reported in Table 7 somewhat underestimate the true magnitude of these correlations but support general inferences about the relative magnitude of these correlations across outcomes.

4.6 Do Teacher Performance Measures Reflect Teacher Effects on Complex Cognitive Skills and Social-emotional Competencies?

Under the Obama administration, the Race to the Top grant competition and state waivers for regulations in the No Child Left Behind Act incentivized states to make sweeping changes to their teacher evaluation systems. Today, most states have implemented new teacher evaluation systems that incorporate multiple measures (Steinberg and Donaldson 2015). Teachers' evaluation ratings are typically derived from a weighted combination of classroom observation scores, assessments of professional conduct, measures of student learning and student surveys. Classroom observations nearly always account for the largest percentage of the overall score, although the weights assigned to measures varies meaningfully across districts and states (Steinberg and Kraft 2016).

The MET Project provides a unique opportunity to further explore the relationship between evaluation metrics used in new teacher evaluation systems and teacher effects on students' complex cognitive skills and social-emotional competencies. In Table 9, I present correlations between the teacher effects I estimate above and a range of evaluation measures from both the same year and prior year. Estimating these relationships using evaluation measures from the prior year serves to eliminate potential upward bias due to correlated errors from a common student sample as described above. I utilize evaluation ratings on two widely used classroom observation instruments: the Framework for Teaching (FFT) and the Classroom Assessment Scoring System (CLASS) (Kane and Staiger 2012). I also include principals' overall

ratings of teachers' performance on a six-point scale and students' opinions of their teachers' instruction captured on the TRIPOD survey (Kane and Cantrell 2010).

I find that neither observation scores, principal ratings, nor student surveys serve as close proxies for teacher effects on the broad set of outcomes. Principal ratings have the strongest relationship with teacher effects on growth mindset with a correlation of .16. Classroom observations scores, particularly on the FFT instrument, are the closest proxy for teacher effects on complex tasks although these correlations are never larger than 0.13. Student surveys have the strongest relationship with teacher effects on students' perseverance and effort in class, although these relationships appear to be largely an artifact of correlated errors as they converge to zero when using estimates based on student ratings from the prior year. These findings suggest that high-stakes decisions based on teacher performance measures commonly used in new evaluation systems largely fail to capture the degree to which teachers are developing students' complex cognitive skills and social-emotional competencies.

5. Robustness Tests

5.1 Teacher Effects – Average Class Residual Estimates

As a robustness check for my preferred model-based maximum likelihood estimation approach, I also estimate the variance of teacher effects by averaging upper and lower bound estimates derived from a two-step estimation approach following Kane et al. (2013). This allows me to relax the random effects normality assumption necessary for equation (3). Given that teacher fixed effects are perfectly collinear with classroom-level controls in the analytic sample, I first fit the covariate-adjustment model described in equation (3), omitting teacher random effects. In a second step, I average student residuals at the teacher level, $\bar{\varepsilon}_{ij}$, to estimate teacher

effects. The variance of these average classroom residuals provide an upper bound estimate. I then shrink the average classroom residuals as described in equation (1). Following Jacob and Lefgren (2008), I estimate λ_j using sample analogs where σ_τ^2 is approximated by subtracting the average of the squared standard errors of the average classroom residuals from the variance of these average classroom residuals ($\hat{\sigma}^2_{\bar{\varepsilon}_{ij}} - \overline{SE^2_{\bar{\varepsilon}_{ij}}}$) and $\sigma^2_{\varepsilon_j}$ is the squared standard error of teacher j's average classroom residuals ($SE^2_{\bar{\varepsilon}_{ij}}$).[10] The variance of these shrunken EB estimates provides a lower-bound estimate. Finally, I average the upper and lower bound estimates to approximate the true teacher variance.

$$\sigma_\tau^2 \approx \frac{(\hat{\sigma}^2_{\tau_{FE}} + \hat{\sigma}^2_{\tau_{EB}})}{2} \quad (5)$$

Two broad findings emerge from comparing alternative estimates in Table 10 to the preferred ML results in Table 6. First, the relative magnitude of teacher effects across outcomes remains similar to ML estimates across model specifications. Second, the magnitudes of the alternative results are slightly smaller than my ML results. This attenuation is largely a mechanical product of the two-stage estimation approach. ML variance estimates are derived from models that include peer controls and teacher random effects simultaneously. In the two-stage process of estimating average class residuals, I first estimate peer effects and then use only the remaining residual variation to quantify teacher effects. If peer effects and teacher effects are correlated, this two-stage approach will cause some variation attributable to teachers to be removed via peer controls in the first stage.

5.2 Removing Prior Test Scores

My identification strategy relies on the random assignment of classroom rosters to

[10] We calculate standard error as the standard deviation of student residuals in a teacher's classroom divided by the square root of the number of students in the teacher's class.

teachers. The inclusion of prior achievement scores from state tests along with additional controls for student and peer characteristics serve to increase the precision of my estimates and to guard against any potential non-random attrition and sorting across classrooms that occurred. The availability of prior state test sores but not prior open-ended tests or social-emotional competencies results in an asymmetry across teacher effects given that some estimates control for lagged outcomes while others do not. I examine the sensitivity of the ML variance estimates from Table 6 and corresponding correlations across teacher effects from Table 7 by comparing them to estimates from models that exclude controls for both prior test scores as well as peer average test scores. These results are presented in Appendix F Table AF1 and AF2.

Comparing results across Tables 6 and AF1 confirms that my primary findings are not a product of the asymmetric set of lagged outcome measures used in these analyses. Unlike prior approaches to estimating teacher effects that rely primarily on lagged test scores to address student sorting, my findings remain consistent when these controls are excluded from the model. Estimates that omit prior scores are slightly larger likely due to an increase in unexplained variance that is then partially attributed to teachers. Results from models that include peer controls increase the most suggesting that the average peer achievement in the prior year plays an important role in capturing peer effects. Correlations among teacher effects are meaningfully larger when models do not include lagged test scores but their relative magnitude across outcomes remains largely the same. Overall, these results suggest the differential findings across outcomes I find are not driven by the inclusion of prior achievement scores from state standardized tests.

5.3 Falsification Tests

At their core, my teacher effect estimates are driven by the magnitude of differences in

classroom means across a range of different outcomes. Given the small number of students taught by each teacher—an average of just over 17 in the analytic sample—it is possible that these estimates are the result of sampling error across classrooms. I test for this by generating a random variable from the standard normal distribution so that it shares the same mean and variance as the outcomes. I then re-estimate my taxonomy of models using these random values as outcomes and report the results in Panel A of Table 11. This test fails to reject the null hypothesis of no teacher effects, demonstrating that my primary estimates are not driven by small sample error.

This randomly generated number test is instructive but does not reflect the patterns of attrition or non-compliance that I observe in the data. An ideal test of bias due to non-random attrition and non-compliance would be to estimate teacher effects on a student characteristic that is correlated with student outcomes, cannot be affected by teachers, and is not included as a covariate in the education production function model. Because such a variable is unavailable, I instead test for teacher effects on a range of student characteristics unaffected by teachers that are included as controls in the models. These characteristics include gender, age, eligibility for free or reduced-price lunch status, and race/ethnicity. I drop a given measure from the set of covariates when I use it as an outcome in these falsification tests. As shown in Table 11 Panel A, I easily reject teacher effects across all of these measures.

In Table 11 Panel B, I further demonstrate that ML estimates are not driven by unexplained variance due to the lower reliability of measures based on constructed response test items or survey questions. Here we, ex post, randomly reassign students to teachers in the analytic sample in a way the exactly replicates the observed number of students with each teacher. This allows me to examine the variance in teacher effects across outcomes when, by

design, teacher effects should be zero. I find no statistically significant teacher effects across all outcomes with the majority of estimates converging to precise zeros. Together, these falsification tests lend strong support to the validity of my teacher effect estimates.

5.4 Potential Reference Bias in Social-Emotional Measures

Previous research has raised concerns about potential reference bias in scales measuring social-emotional skills based on student self-reporting (Duckworth and Yeager 2015). For example, studies have found that over-subscribed urban charter schools with explicit school-wide cultures aimed at strengthening students' social-emotional competencies appear to negatively affect students' self-reported grit, but have large positive effects on achievement and persistence in school (West et al. 2016; Dobbie and Fryer 2013). Notably, West et al. (2016) find little evidence of reference bias on the growth mindset scale, possibly because it asks students about beliefs which are not easily observed and, thus, less likely to be judged in reference to others.

I examine whether students' responses on self-reported measures of grit, growth mindset and effort in class may be subject to reference bias in my sample of traditional public schools in large urban districts. I do this by exploring how the direction and magnitude of the relationship between these social-emotional measures and student achievement gains on state standardized tests change when collapsed from the student-level to the class- and school-levels. Employing this same test, West et al. (2016) find suggestive evidence of reference bias in self-reported measures of grit, conscientiousness and self-control in a sample of students attending traditional, charter and exam schools in Boston. They find that correlations between social-emotional measures and overall student gains become negative when collapsed to the school-level. This is analogous to the classic example of reference bias in cross-cultural surveys where, despite a widely acknowledged cultural emphasis on conscientious behavior, individuals in East Asian

countries rate themselves lower in conscientiousness than do individuals in any other region (Schmitt et al. 2007).

I find no compelling evidence of reference bias at either the class level or the school level in the MET data. As shown in Table 12, simple Pearson correlation coefficients between the four social-emotional measures and student gains on state math and ELA tests are all small, positive and statistically significant at the student level. Collapsing the data at the classroom or school level does not reverse the sign of any of the student-level correlations, and, if anything, increases the positive relationships between self-reported social-emotional competencies and student gains. Although I cannot rule out the potential of reference bias in the measures, it does not appear as though teachers or schools where students are making larger achievement gains are also systematically changing students' perceptions of what constitutes gritty behavior and high levels of effort. Additionally, the MET Project's experimental design limits the identifying variation to within school-grade cells, eliminating any potential for reference bias at the school-level and grade-level within a school.

6. Conclusion

The hallmark education policy reforms of the early 21st century — school accountability and teacher evaluation — created strong incentives for educators to improve student performance on state standardized tests. There is no doubt authentic improvements in students' underlying content knowledge and basic skills assessed on these tests are important for success in school and later in life. As I show using the ELS dataset, standardized test scores are strong predictors of a range of adult outcomes. However, these tests provide a narrow measure of the range of abilities and competencies that predict positive adult outcomes. Questions remain about whether

those teachers and schools that are judged as effective by state standardized tests are also developing students' more complex cognitive skills and social-emotional competencies. My results suggest that this is not always the case.

The large differences in teachers' ability to raise student performance on achievement tests (Chetty, Friedman and Rockoff 2014a; Hanushek and Rivkin 2010) and the inequitable distribution of those teachers who are most successful at raising achievement (Clotfelter, Ladd, Vigdor 2006; Lankford, Loeb, Wyckoff 2002) have become major foci of academic research and education policy. The substantial variation I find in teacher effects on students' complex task performance and social-emotional competencies further reinforces the importance of teacher quality but complicates its definition. Measures of teachers' contribution to their students' performance on state tests in math are strong proxies for their effects on students' ability to solve complex math problems. However, teacher effects on state ELA tests contain more limited information about how well a teacher is developing students' abilities to reason about and draw inferences from texts. Teacher effects on state tests are even weaker indicators of the degree to which teachers are developing students' social-emotional competencies. Even teachers who excel at developing competencies such as grit are not consistently the same as those that develop other competencies such as growth mindset.

Teaching core academic skills along with social-emotional competencies and the ability to perform unstructured tasks should not be viewed as competing priorities in a zero sum game. Elevating the importance of these new foundational skills does not require schools to make tradeoffs such as deciding between expanding instructional time in core subjects or teaching the arts and foreign languages. The MET data suggest that there are teachers who teach core academic subjects in ways that also develop students' complex problem-solving skills and social-

emotional competencies. We need to know what instructional practices allow these teachers to develop a wider range of students' skills and competencies than are commonly assessed on state achievement tests.

Current accountability and evaluation systems in education provide limited incentives for teachers to focus on helping students develop complex problem-solving skills and social-emotional competencies. Findings from this paper suggest that neither observation scores, principal ratings, nor student surveys are serving as close proxies for teacher effects on these skills and competencies. New computer-adaptive assessments aligned with the Common Core State Standards move in the direction of assessing more complex cognitive skills (Doorey and Polikoff, 2016) but are facing growing opposition. A move towards complementing or replacing states tests with assessments of more complex cognitive skills would help better align incentives for teachers but faces important challenges given the traditionally lower reliability and higher cost of creating and scoring constructed response items and the possible public resistance to tests that label fewer students as proficient.

Developing practical and reliable measures of students' social-emotional competencies that could be used in school accountability or teacher evaluation systems poses an even greater challenge. Psychologists have argued that the social-emotional measures used in this study are not sufficiently robust to be used in high-stakes settings to compare teachers across schools (Duckworth and Yeager 2015). Student self-reports or teacher assessments of social-emotional measures are easy to game, and we know little about their properties when stakes are attached. There exists real potential to improve the reliability and robustness of these measures, but it may be that observable student outcomes such as attendance and disciplinary incidents are ultimately more tractable measures for policy purposes (Whitehurst, 2015). As Einstein observed,

"Everything that counts cannot necessarily be counted." What is clear is that our current conception of teacher effectiveness needs to be expanded to encompass the multiple ways in which teachers affect students' success in school and life.

References

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim D. Kautz, "Personality Psychology and Economics," NBER Working Paper No. w16822, National Bureau of Economic Research, (2011).

Bill & Melinda Gates Foundation. User Guide to Measures of Effective Teaching Longitudinal Database (MET LDB). Inter-University Consortium for Political and Social Research, (2013). http://www.icpsr.umich.edu/icpsrweb/METLDB/studies/34771

Blackwell, Lisa S., Kali H. Trzesniewski, and Carol Sorich Dweck, "Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention," *Child Development*, 78 (2007), 246-263.

Blazar, David.,Matthew A. Kraft, "Teacher and Teaching Effects on Students' Attitudes and Behaviors." *Educational Evaluation and Policy Analysis, 39* (2017), 146-170.

Bowles, S., Gintis, H., and Osborne, M. "The Determinants of Earnings: A Behavioral Approach." *Journal of Economic Literature*, *39*(2001), 137-176.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas Ter Weel, "The Economics and Psychology of Personality Traits," *Journal of Human Resources*, 43 (2008), 972-1059.

Bowen, William G., Matthew M. Chingos, and Michael S. McPherson, *Crossing the Finish Line: Completing College at America's Public Universities* (Princeton, NJ: Princeton University Press, 2009).

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126 (2011), 1593-1660.

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104 (2014b), 2633-2679.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources*, 41 (2006), 778-820.

Corcoran, Sean. P., Jennifer L. Jennings, and Andrew A. Beveridge, "Teacher Effectiveness on High-and Low-Stakes Tests." New York University Working Paper, (2012).

http://www.nyu.edu/projects/corcoran/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf

Dee, Thomas S., and James Wyckoff, "Incentives, Selection, and Teacher Performance: Evidence from IMPACT," *Journal of Policy Analysis and Management*, 34 (2015), 267-297.

Deming, David J., "The Growing Importance of Social Skills in the Labor Market," NBER Working Paper No. w21473, National Bureau of Economic Research, (2015).

Dobbie, Will, and Roland G. Fryer Jr., "The Medium-Term Impacts of High-Achieving Charter Schools on Non-Test Score Outcomes," NBER Working Paper No. w19581, National Bureau of Economic Research, (2013).

Doorey, Nancy, and Morgan Polikoff. "Evaluating the Content and Quality of Next Generation Assessments." *Thomas B. Fordham Institute* (2016).

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly, "Grit: Perseverance and Passion for Long-Term Goals," *Journal of Personality and Social Psychology*, 92 (2007), 1087-1101.

Duckworth, Angela Lee, and Patrick D. Quinn, "Development and Validation of the Short Grit Scale (GRIT–S)," *Journal of Personality Assessment*, 91 (2009), 166-174.

Duckworth, A. L., Grant, H., Loew, B., Oettingen, G., & Gollwitzer, P. M. "Self-regulation Strategies Improve Self-discipline in Adolescents: Benefits of Mental Contrasting and Implementation Intentions." *Educational Psychology*, *31* (2011), 17-26.

Duckworth, A. L., Kirby, T. A., Gollwitzer, A., & Oettingen, G. From Fantasy to Action: Mental Contrasting with Implementation Intentions (MCII) Improves Academic Performance in Children. *Social Psychological and Personality Science*, *4* (2013), 745-753.

Duckworth, Angela L., and David Scott Yeager, "Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes," *Educational Researcher*, 44 (2015), 237-251.

Dweck, Carol, *Mindset: The New Psychology of Success*, (Random House, 2006).

Eskreis-Winkler, Lauren, Elizabeth P. Shulman, Scott A. Beal, and Angela L. Duckworth, "The Grit Effect: Predicting Retention in the Military, the Workplace, School and Marriage," *Frontiers in Psychology*, Feb (2014), 5-36.

Gershenson, Seth, "Linking Teacher Quality, Student Attendance, and Student Achievement," *Education Finance and Policy*, 11 (2016), 125-149.

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge. “Can Value-Added Measures of Teacher Performance be Trusted?." *Education Finance and Policy, 10:1* (2015) 117-156.

Hanushek, Eric A., and Steven G. Rivkin, "Generalizations about Using Value-Added Measures of Teacher Quality,” *The American Economic Review*, 100 (2010), 267-271.

Heckman, James J., Rodrigo Pinto, and Peter A. Savelyev, “Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes,” NBER Working Paper No. w18581, National Bureau of Economic Research, (2012).

Heckman, James J., Jora Stixrud, and Sergio Urzua, “The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior,” *Journal of Labor Economics*, 24 (2006), 411-482.

Jackson, C. Kirabo. “What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes.” No. w22226. National Bureau of Economic Research, (2016).

Jacob, Brian and Lars Lefgren, “Principals as Agents: Subjective Performance Assessment in Education,” *Journal of Labor Economics*, 26 (2008), 101-136.

Jennings, Jennifer L., and Thomas A. DiPrete, “Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology of Education*, 83 (2010), 135-159.

Kalogrides, Demetra, and Susanna Loeb, “Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools," *Educational Researcher*, 42 (2013), 304-316.

Kane, Thomas J., and Steve Cantrell. “Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.” MET Project Research Paper, Bill & Melinda Gates Foundation, (2010).

Kane, Thomas J., Daniel F. McCaffrey, Tre Miller, and Douglas O. Staiger, “Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment,” *Seattle, WA: Bill and Melinda Gates Foundation*, (2013).

Kane, Thomas J., and Douglas O. Staiger, “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation,” NBER Working Paper No. w14607, (2008).

Kane, Thomas J., and Douglas O. Staiger, “Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.” MET Project. *Bill & Melinda Gates Foundation*. (2012).

Koedel, Cory, “Teacher Quality and Dropout Outcomes in a Large, Urban School District,” *Journal of Urban Economics*, 64 (2008), 560-572.

Koedel, Cory, and Julian R. Betts, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," *Education Finance and Policy*, 6 (2011), 18-42.

Ladd, Helen F. and Lucy C. Sorensen. "Returns to Teacher Experience: Student Achievement and Motivation in Middle School." *Education Finance and Policy*. (2017).

Lankford, Hamilton, Susanna Loeb, and James Wyckoff, "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis*, 24 (2002), 37-62.

Le, Vi-Nhuan, Brian M. Stecher, J. R. Lockwood, Laura S. Hamilton, and Abby Robyn, *Improving Mathematics and Science Education: A Longitudinal Investigation of the Relationship between Reform-Oriented Instruction and Student Achievement* (Rand Corporation, 2006).

Lu, Qian, "The End of Polarization? Technological Change and Employment in the US Labor Market," Working Paper, (2015).

Miu, A. S., & Yeager, D. S. "Preventing Symptoms of Depression by Teaching Adolescents that People Can Change Effects of a Brief Incremental Theory of Personality Intervention at 9-month Follow-up." *Clinical Psychological Science*, *3* (2015), 726-743.

MET Project, "Memorandum: MET Test Recommendations" October 8th, (2009).

Moffitt, Terrie E., Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, HonaLee Harrington, Renate Houts et al., "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety," *Proceedings of the National Academy of Sciences*, 108 (2011), 2693-2698.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges, "How Large are Teacher Effects?," *Educational Evaluation and Policy Analysis*, 26 (2004), 237-257.

National Research Council. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. J.W. Pellegrino and M.L. Hilton, Editors. (The National Academies Press , 2012).

OECD (2013) PISA 2015: Draft Collaborative Problem Solving Framework. http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. "Mind-set Interventions are a Scalable Treatment for Academic Underachievement." *Psychological Science*, 26 (2015), 784–793.

Raudenbush, Stephen W., and Anthony S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Vol. 1. (Sage, 2002).

Rockoff, Jonah E., "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *The American Economic Review*, 94 (2004), 247-252.

Rothstein, Jesse, "Teacher Quality in Educational Production: Tracking, Decay and Student Achievement," *Quarterly Journal of Economics*, 125 (2010).

Ruzek, Erik A., Thurston Domina, AnneMarie M. Conley, Greg J. Duncan, and Stuart A. Karabenick, "Using value-added models to measure teacher effects on students' motivation and achievement," *The Journal of Early Adolescence*, (2014), 1-31.

Schmitt, David P., Jüri Allik, Robert R. McCrae, and Verónica Benet-Martínez, "The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 nations," *Journal of Cross-Cultural Psychology*, 38 (2007), 173-212.

Shechtman, Nicole, Angela H. DeBarger, Carolyn Dornsife, Soren Rosier, and Louise Yarnall. "Promoting Grit, Tenacity, and Perseverance: Critical Factors for Success in the 21st Century." *Washington, DC: US Department of Education, Department of Educational Technology* (2013): 1-107.

Snyder, Thomas D., and Sally A. Dillow. *Digest of Education Statistics 2013*. (National Center for Education Statistics, 2015).

Spearman, Charles, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, 15 (1904), 72-101.

Steinberg, Matthew P., and Matthew A. Kraft. "The Sensitivity of Teacher Performance Ratings to the Design of Teacher Evaluation Systems." (2016). Working Paper access at http://scholar.harvard.edu/mkraft/publications/sensitivity-teacher-performance-ratings-design-teacher-evaluation-systems

Sule, Alan, Teodora Boneva, and Seda Ertac. "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit. (2016). Working Paper accessed at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2761390

Thompson, Paul N., Cassandra M. Guarino, and Jeffrey M. Wooldridge. "An Evaluation of Teacher Value-Added Models with Peer Effects." Working Paper, (2015).

Todd, Petra E., and Kenneth I. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal*, 113 (2003), F3-F33.

Tough, Paul, *How Children Succeed*, (Random House, 2013).

West, Martin R., Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher F.O. Gabrieli, and John D.E. Gabrieli, "Promise and Paradox Measuring Students' Non-Cognitive Skills and the Impact of Schooling," *Educational Evaluation and Policy Analysis*, (2016). 148-170.

Weinberger, Catherine J., "The Increasing Complementarity between Cognitive and Social Skills," *Review of Economics and Statistics*, 96 (2014), 849-861.

Whitehurst, Grover J. "Hard Thinking on Soft Skills." *Evidence Speaks Reports*, Brookings Institute, 1, no. 14 (2016): 5.

Yeager, David S., Dave Paunesku, Gregory M. Walton, and Carol S. Dweck. "How Can We Instill Productive Mindsets at Scale? A Review of the Evidence and an Initial R&D Agenda." In *white paper prepared for the White House meeting on "Excellence in Education: The Importance of Academic Mindsets," 2013. http://homepage. psy. utexas. edu/HomePage/Group/YeagerLAB/ADRG/Pdfs/Yeager et al R&D agenda-6-10-13. pdf*

Yeager, D. S., Johnson, R., Spitzer, B. J., Trzesniewski, K. H., Powers, J., & Dweck, C. S. "The Far-reaching Effects of Believing People Can Change: Implicit Theories of Personality Shape Stress, Health, and Achievement During Adolescence." *Journal of personality and social psychology*, *106* (2014), 867.

Yeager, D. S., Miu, A. S., Powers, J., & Dweck, C. S. Implicit Theories of Personality and Attributions of Hostile Intent: A Meta-analysis, an Experiment, and a Longitudinal Intervention. *Child development*, *84* (2013), 1651-1667.

Yeager, D. S., Walton, G. M., Brady, S. T., Akcinar, E. N., Paunesku, D., Keane, L., ... & Gomez, E. M. "Teaching a Lay Theory Before College Narrows Achievement Aaps at Scale". *Proceedings of the National Academy of Sciences*, *113* (2016), E3341-E3348.

Yuan, Kun, and V. Le, "Estimating the Percentage of Students Who Were Tested on Cognitively Demanding Items through the State Achievement Tests," (Santa Monica, CA: RAND Corporation, 2012).

Yuan, Kun, and Vi-Nhuan Le. "Measuring Deeper Learning through Cognitively Demanding Test Items: Results from the Analysis of Six National and International Exams. Research Report." *RAND Corporation* (2014).

Zeiser, Kristina .L., James Taylor, Jordan Rickles, Michael S. Garet, Michael Segeritz, *Evidence of Deeper Learning. Findings from the Study of Deeper Learning: Opportunities and Outcomes*, (American Institutes for Research, 2014). http://www.air.org/sites/default/files/downloads/report/Report_3_Evidence_of_Deeper_Learning_Outcomes.pdf

Figures

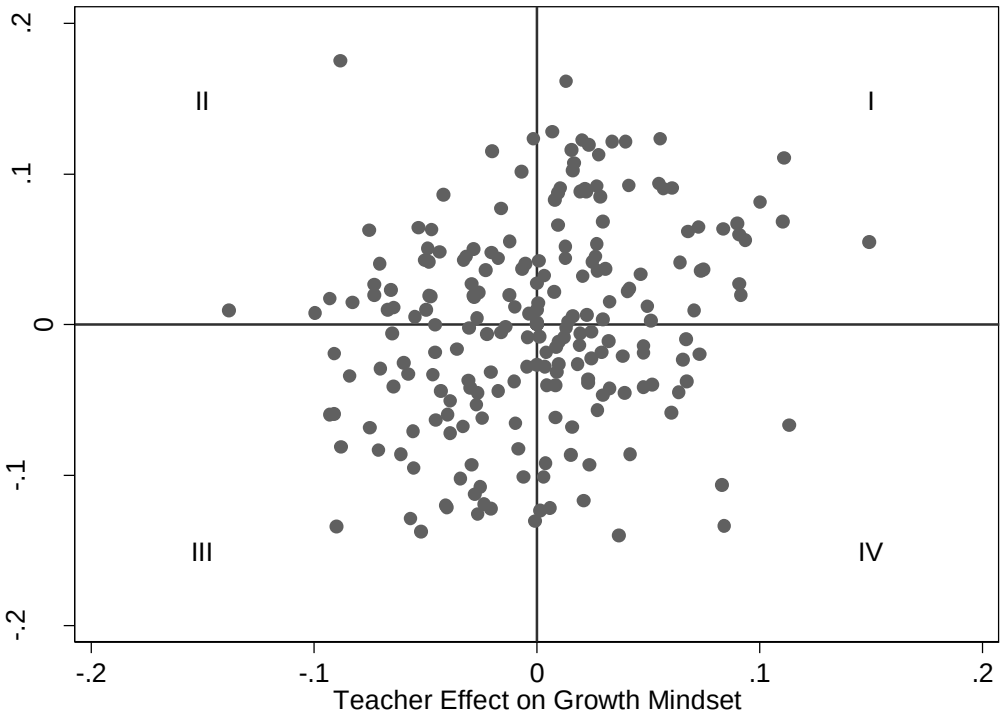

Figure 1: Scatterplot of teacher effects on state math test and growth mindset.

Figure 2: Re-estimated correlations coefficients from Table 7 using restricted samples of successively larger minimum class size requirements

Tables

Table 1: Student & Teacher Characteristics

	Students			Teachers	
	Study	U.S. Public Schools in Cities	U.S. Public Schools	Study	U.S. Public Schools
Age	9.50				
Gifted Status	0.07		0.06		
Special Education Status	0.08		0.13		
English Language Learner	0.15	0.14	0.10		
Free or Reduced Price Lunch	0.62		0.52		
Male	0.49		0.51	0.08	0.24
Asian	0.08	0.07	0.05		
White	0.24	0.30	0.49	0.62	0.82
African American	0.36	0.25	0.16	0.33	0.07
Hispanic	0.29	0.35	0.26	0.05	0.08
1 Year of Experience in District				0.07	0.09*
2-3 Years of Experience in District				0.18	
4-6 Years of Experience in District				0.23	0.33†
7-10 Years of Experience in District				0.24	
11-20 Years of Experience in District				0.29	0.36
> 20 Years of Experience in District				0.12	0.21
Graduate Degree				0.50	0.56
n	4092	14,457,000	50,132,000	236	3,119,001

Notes: The study sample consists of all 4th and 5th grade students taught by general education teachers who participated in the randomization study with valid data for student demographics and at least one academic or social-emotional outcome, as well as prior test scores on both math and ELA state exams. Sources for U.S. public school student and teacher data is the NCES Digest of Education Statistics and Census CPS on School Enrollment for male percentage. Data for all U.S. public schools is from 2013/14. Data for U.S. public schools in cities is from 2011/12.

* Corresponds to less than 3 years of experience

† Corresponds to 3-9 years of experience

Table 2: The Predictive Validity of Self-Reported Character Skills on Education, Employment, Personal, and Civic Outcomes

	Education	Labor Market		Personal		Civic	
	Bachelor's Degree	Employed	Employment Income	Teen Parent	Married	Voted in Presidential Election	Volunteered
Academic Achievement	0.156***	0.033***	3125.511***	-0.027***	0.005	0.070***	0.073***
	(0.006)	(0.007)	(341.105)	(0.004)	(0.007)	(0.007)	(0.007)
Grit: Perseverance of Effort	0.058***	0.026***	1631.608***	-0.008*	0.019**	0.035***	0.036***
	(0.006)	(0.006)	(313.679)	(0.003)	(0.006)	(0.006)	(0.006)
Growth Mindset in Math	0.011*	0.006	848.157**	-0.006*	-0.009	0.019**	0.008
	(0.005)	(0.006)	(324.151)	(0.003)	(0.006)	(0.006)	(0.006)
N	8647	8643	8647	8248	8566	8542	8567
R-squared	0.209	0.012	0.042	0.035	0.002	0.045	0.046

Notes: * p<0.05, ** p<0.01, *** p<0.001. Academic Achievement is the average of scores on math and reading tests. Measures of grit and growth mindest are proxy measures constructed from questions available in the ELS dataset. All models include controls for students' gender and race as well as parental level of education and household income. Employment income is a self-reported measure of all earnings (in dollars) before taxes and deductions in 2011.

Table 3: Disattenuated Correlations among State Tests, Complex Tasks and Social-Emotional Measures

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	0.81	1.00					
BAM Math	0.81	0.73	1.00				
SAT9-OE Reading	0.49	0.56	0.69	1.00			
Growth Mindset	0.29	0.35	0.35	0.28	1.00		
Grit: Consistency	0.35	0.40	0.36	0.31	0.46	1.00	
Grit: Perseverence	0.23	0.25	0.21	0.22	0.07	0.33	1.00
Effort in Class	0.40	0.41	0.38	0.35	0.26	0.59	0.91

Notes: n=5610. All correlations are statistically significant at the p<.01 level, except for the correlation between Growth Mindset and Grit: Perseverance, which is statistically significant at the p<.05 level.

Table 4: Testing Post-Attrition Randomization Balance in Student Demographic and Prior Achievement across Teachers in the Same Randomization Block

	Randomization Teacher	
	F-Statistic	P-value
Male	0.241	1.000
Age	0.763	0.997
Gifted Status	1.460	0.000
Special Education Status	0.957	0.668
English Language Learner	1.762	0.000
Free or Reduced Price Lunch	0.559	1.000
White	0.383	1.000
African American	0.588	1.000
Hispanic	0.633	1.000
Asian	0.620	1.000
State Math 2010	1.013	0.433
State ELA 2010	1.071	0.222
n	4092	

Notes: F-Statistics and corresponding p-values are from joint tests of teacher fixed effects from a model where a given student characteristic, demeaned within randomization blocks, is regressed on teacher fixed effects.

Table 5: The Relationship between Student Characteristics and Randomly Assigned Teacher Characteristics Post-Attrition

	Teacher Value-Added in Prior Year			
	State Math	State ELA	BAM	SAT9-OE
Male	-0.001	0.000	-0.003	0.001
	(0.004)	(0.003)	(0.003)	(0.003)
Age	0.001	0.002	0.008	-0.007
	(0.006)	(0.005)	(0.006)	(0.008)
Gifted Status	0.035	0.002	0.003	-0.015
	(0.033)	(0.022)	(0.021)	(0.020)
Special Education Status	0.011	0.003	0.015	0.005
	(0.008)	(0.008)	(0.012)	(0.020)
English Language Learner	-0.018	-0.014	-0.005	-0.013
	(0.009)	(0.010)	(0.012)	(0.014)
Free or Reduced Price Lunch	0.017*	0.001	0.001	0.011
	(0.008)	(0.006)	(0.008)	(0.012)
White	-0.010	-0.009	-0.003	-0.013
	(0.005)	(0.006)	(0.004)	(0.008)
African American	0.011	0.005	-0.004	0.013
	(0.006)	(0.007)	(0.006)	(0.010)
Hispanic	-0.009	-0.006	-0.001	0.000
	(0.005)	(0.006)	(0.006)	(0.009)
Asian	0.010	0.014	0.011	0.001
	(0.007)	(0.011)	(0.006)	(0.009)
State Math 2010 (z-scores)	0.005	0.003	0.003	-0.003
	(0.004)	(0.003)	(0.003)	(0.004)
State ELA 2010 (z-scores)	0.005	0.004	0.001	-0.003
	(0.003)	(0.003)	(0.003)	(0.004)
n	4092	4041	4076	4041

Notes: *p<0.05. Each cell presents results from a separate regression of the value added estimate for the teacher students were randomly assigned to by MET Project researchers on a given student characteristic. Value-added estimates are in student standard deviation units (Math .226; ELA .170; BAM .211; SAT9-OE .255).

Table 6: Model-based Restricted Maximium Likelihood Estimates of Teacher Effects on State Tests, Complex Tasks and Social-Emotional Measures

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
State Math	4,075	0.157***	0.168***	0.136***	0.152***	0.122***
State ELA	4,074	0.140***	0.172***	0.143***	0.148***	0.125***
BAM Math	3,746	0.137***	0.168***	0.129***	0.150***	0.112**
SAT9-OE Reading	3,766	0.163***	0.178***	0.162***	0.175***	0.158***
Growth Mindset	3,551	0.201***	0.156**	0.138*	0.168***	0.157**
Grit: Consistency	3,473	0.088	0.088	0.074	0.098	0.100
Grit: Perseverance	3,473	0.153**	0.153**	0.140*	0.152**	0.141*
Effort in Class	3,435	0.158***	0.157**	0.172***	0.113*	0.140*
Survey-based Controls			Yes	Yes	Yes	Yes
Peer-level Controls		Yes		Yes		Yes
School FE		Yes				
Randomization Block FE			Yes	Yes	Yes	Yes

Notes: * p<0.05, ** p<0.01, *** p<0.001. Cells report the standard deviation of teacher effect estimates from separate regressions. All models include controls for student's gender, age, race, FRPL, English proficiency status, special education status, and participation in a gifted and talented program. Survey-based controls include self-reported prior grades, the number of books at home, the degree to which English is spoken at home, and the number of computers at home. Peer-level controls are classroom averages of prior achievement as well as all administrative and survey-based measures described above.

Table 7: Correlations of Teacher Effects on State Tests, Complex Tasks, and Social-Emotional Measures

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	0.60***	1.00					
BAM Math	0.65***	0.36***	1.00				
SAT9-OE Reading	0.35***	0.25***	0.43***	1.00			
Growth Mindset	0.22***	0.19**	0.12	0.22***	1.00		
Grit: Consistency	0.17**	0.20**	0.10	-0.02	0.22***	1.00	
Grit: Perseverence	-0.06	-0.02	0.10	0.18**	-0.01	0.03	1.00
Effort in Class	0.06	0.08	0.14*	0.09	-0.06	0.06	0.61***

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. Classroom effects are derived using the model reported in Column 3 of Table 6.

Table 8: Teacher Effect Correlations Within and Between Sections

	n (Teachers)	Correlation Within Sections	Correlation Between Sections
State Math & BAM Math	433	0.37	0.15
State Math & Effort in Class	433	0.20	0.15
BAM Math & Effort in Class	433	0.18	0.20
State ELA & SAT9-OE Reading	497	0.28	0.15
State ELA & Effort in Class	497	0.15	0.04
SAT9-OE Reading & Effort in Class	497	0.17	0.13

Notes: Estimates derived using value-added estimates provided by the MET Project for subject-specialist teachers in Year 1 of the study.

Table 9: Correlations of Teacher Performance Measures with Teacher Effects on State Tests, Complex Tasks, and Social-Emotional Measures

	FFT		CLASS		Student Surveys		Principal Ratings
	Current Year	Prior Year	Current Year	Prior Year	Current Year	Prior Year	Current Year
State Math	0.083	-0.005	0.070	0.029	0.000	0.036	0.173*
State ELA	0.105	-0.002	0.037	0.079	0.066	0.123	0.108
BAM Math	0.117	0.056	0.057	0.042	0.116	0.104	0.095
SAT9-OE Reading	0.124	0.023	0.077	0.064	0.038	0.069	0.037
Growth Mindset	0.103	0.064	0.111	0.075	0.010	0.092	0.155*
Grit: Consistency	0.041	0.011	0.026	0.054	0.080	-0.011	0.039
Grit: Perseverence	0.079	-0.030	0.069	0.034	0.190**	-0.035	-0.127
Effort in Class	0.125	-0.020	0.126	0.058	0.192**	0.005	-0.087

Notes: *p<0.05; **p<0.01; ***p<0.001. Classroom effects are derived using the model reported in Column 3 of Table 6. n ranges from 191 (principal ratings) to 235 (FFT & CLASS). FFT and CLASS scores are calculated using the first factor from a Principal Component Analysis of the average domain-level scores for each instrument.

Table 10: Average of Shrunken and Unshrunken of Teacher Effects on State Tests, Complex Tasks and Social-Emotional Measures

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
State Math	4,075	0.140	0.125	0.091	0.100	0.114
State ELA	4,074	0.120	0.123	0.088	0.082	0.070
BAM Math	3,744	0.134	0.130	0.095	0.086	0.135
SAT9-OE Reading	3,766	0.160	0.146	0.114	0.119	0.101
Growth Mindset	3,551	0.214	0.157	0.107	0.135	0.107
Grit: Consistency	3,473	0.124	0.105	0.088	0.102	0.095
Grit: Perseverance	3,473	0.164	0.143	0.107	0.113	0.103
Effort in Class	3,435	0.177	0.151	0.121	0.138	0.166
Survey-based Controls			Yes	Yes	Yes	Yes
Peer-level Controls		Yes		Yes		Yes
School FE		Yes				
Randomization Block FE			Yes	Yes	Yes	Yes

Notes: Cells represent estimates from separate regressions. Statistical significance not calculated given estimates represent the average across shrunken and unshrunken estimates. See notes from Table 6 for further model details.

Table 11: Falification Tests of Teacher Effects

	n	(1)	(2)	(3)	(4)	(5)
		Panel A: Actual Students / Immutable Outcome				
		Actual Teacher			Randomly Assigned	
Random Number	4,092	0.000	0.000	0.000	0.000	0.000
Male	4,092	0.000	0.000	0.000	0.000	0.000
Age	4,092	0.049	0.046	0.048	0.043	0.040
Free or Reduced Price Lunch	2,326	0.000	0.000	0.000	0.000	0.000
White	4,092	0.000	0.000	0.000	0.000	0.000
African American	4,092	0.030	0.022	0.025	0.000	0.000
Hispanic	4,092	0.028	0.032	0.029	0.029	0.024
		Panel B: Re-randomized Students to Teachers				
State Math	4,075	0	0.024	0.016	-	-
State ELA	4,074	0	0.035	0.015	-	-
BAM Math	3,723	0	0	0	-	-
SAT9-OE Reading	3,753	0	0	0	-	-
Growth Mindset	3,547	0	0	0	-	-
Grit: Consistency	3,463	0.082	0.082	0.081	-	-
Grit: Perseverance	3,463	0	0	0	-	-
Effort in Class	3,435	0	0	0	-	-
Survey-based Controls			Yes	Yes	Yes	Yes
Peer-level Controls		Yes		Yes		Yes
School FE		Yes				
Randomization Block FE			Yes	Yes	Yes	Yes

Notes: Cell represent model-based restricted maximium likelihood estimates from separate regressions. No estimates are statistically significant. The sample size for Free or Reduced Price Lunch is limited because one participating district did not provide this information. All samples restricted to require at least 5 students per teacher. See notes from Table 6 for further model details.

Table 12: Student, Class, and School Level Correlations between Social-Emotional measures and Gain Scores on State Tests

	State Math Gains			State ELA Gains		
	Student-level	Class-level	School-level	Student-level	Class-level	School-level
Growth Mindset	0.06**	0.23**	0.08	0.10***	0.25**	0.30**
Grit: Consistency	0.08***	0.19**	0.10	0.12***	0.26***	0.13
Grit: Perseverance	0.05**	0.08	0.19*	0.08***	0.15*	0.17*
Effort in Class	0.11***	0.24**	0.43***	0.10***	0.27***	0.29**
n students	4799	266	149	4799	266	149

Notes: *p<0.05; **p<0.01; ***p<0.001. Test scores gains are the residuals from regressions of a student's current score on cubic functions of their prior math and ELA state test scores. Reported sample sizes represent the largest sample among the four social-emotional measures.

Appendix A

MET Short Grit Scale

Elementary Items:

1. I often set a goal but later choose to pursue a different one.* (CoI)
2. Sometimes, when I'm working on a project, I get distracted by a new and different [topic].* (CoI)
3. I have been obsessed with a certain idea or project for a short time but later I [lose that interest].* (CoI)
4. It's hard for me to finish projects that take a long time to complete.* (CoI)
5. I finish whatever I begin. (PoE)
6. If something is hard to do and I begin to fail at it, I keep trying anyway. (PoE)
7. I am a hard worker. (PoE)
8. I try to do a good job on everything I do. (PoE)

CoI = Items that comprise the Consistency of Interest subscale
PoE = Items that comprise the Perseverance of Effort subscale

* Items are reverse coded

Response scale:

Not like me at all (1)
Not much like me (2)
Some-what like me (3)
Mostly like me (4)
Very much like me (5)

MET Growth Mindset Scale

Elementary & Secondary Items:

1. Your intelligence is something you can't change very much.*
2. You have a certain amount of intelligence, and you can't really do much to change [that].*
3. You can learn new things, but you can't really change your basic intelligence.*

* Items are reverse coded

Response Scale:

Disagree A Lot (1)

Disagree (2)
Disagree A Little (3)
Agree a Little (4)
Agree (5)
Agree a Lot (6)

MET TRIPOD items used to measure Effort in Class

Elementary & Secondary Items:

1. I have done my best quality work in this class.
2. I have pushed myself hard to understand my lessons in this class.
3. When doing schoolwork in this class, I try to learn as much as I can and I don't worry how long it takes.
4. In this class I stop trying when the work gets hard.
5. In this class I take it easy and do not try very hard to do my best.
6. When homework is assigned for this class, how much do you usually complete?

Response scale for items 1-5:

Totally Untrue (1)
Mostly Untrue (2)
Somewhat (3)
Mostly True (4)
Totally True. (5)

Response scale for item 6:

Never Assigned (1)
None of it (2)
Some of it (3)
Most of it (4)
All (5)
All plus some extra (6)

Appendix B

Measures used in the Educational Longitudinal Study analyses

Social-emotional Measures
All questions were asked using a 1-4 Likert Scale, with "Strongly Disagree", "Disagree", "Agree" and "Strongly Agree" being assigned values 1 through 4, respectively. For both variables, indices were created by averaging the responses to all sub-questions identified as pertaining to effort and growth mindset from the survey. These questions were as follows:

Growth mindset (in math) (Taken from ELS 2002 Student Questionnaire, Question 88):

a) Most people can learn to be good at math
b) You have to be born with the ability to be good at math (reverse coded)

Grit: Perseverance of Effort (Taken from ELS 2002 Student Questionnaire, Question 89):

a) When studying, I try to work as hard as possible
b) When studying, I keep working even if the material is difficult
c) When studying, I try to do my best to acquire the knowledge and skills taught
d) When Studying, I put forth my best effort

Achievement Measures

Input variables, including a composite of math and reading test scores and constructed scores for growth mindset and effort, were taken from the original ELS 2002 base year survey. Math and reading assessments were conducted by the ELS group, using materials adapted from previous studies. Math tests included questions on arithmetic, algebra, geometry, statistics, and other advanced material. Reading tests included comprehension questions on passages from literary, science, and social science material. Both tests were predominantly multiple-choice, although the math test did include a few open ended questions which were scored without partial credit. For both tests, all students took a short "first-stage" test, and then were scored and assigned to a "second-stage" test based on their previous performance. This was done to allow for increased accuracy of the results given the short window of testing time and avoid ceiling and floor effects. Test scores for both reading and math are given in the dataset as standardized Z-scores, which were then averaged and re-standardized to create the "average score" variable used in this analysis. This variable has a mean of zero and a standard deviation of one.

Adult Outcome Measures

Outcome variables were taken from follow-up data collected by the ELS in 2012. Outcome variables were treated to ensure that missing values were dropped in each relevant regression. Outcomes are further defined below:

- Bachelor's Degree: Coded as 1 if respondent reported receiving a Bachelor's Degree by the 2012 follow-up survey, 0 if they reported receiving any amount of education less than a Bachelor's Degree.
- Employed: Coded as 1 if respondent reported having one or more (at least part-time) jobs, 0 for those who did not work.
- Employment Income: Self-reported annual income from employment.

- Married: Coded as 1 for all married respondents, 0 for all other domestic arrangements.
- Teen Parent: Coded as 1 for respondents who reported first having a child before or at the age of 19, 0 for respondents who reported having a child after age 19. All childless respondents were dropped.
- Registered to Vote: Coded as 1 for respondents who reported being currently registered to vote, 0 if not registered.
- Voted in Presidential Election: Coded at 1 for respondents who reported voting in the 2008 presidential election, 0 if they did not vote.
- Volunteered: Coded as 1 for respondents who reported having performed unpaid volunteer work in the past two years, 0 for those who did not.

Appendix C

I arrive at estimates for Table 3 by disattenuating the raw correlation coefficients in Table AC1 below using the Spearman (1904) adjustment.

Table AC1: Correlations among State Tests, Complex Tasks and Social-Emotional Measures

	State Math	State ELA	BAM Math	SAT9-OE Reading	Growth Mindset	Grit: Consistency	Grit: Perseverance
State Math	1.00						
State ELA	0.74	1.00					
BAM Math	0.66	0.58	1.00				
SAT9-OE Reading	0.43	0.49	0.54	1.00			
Growth Mindset	0.25	0.29	0.26	0.23	1.00		
Grit: Consistency	0.27	0.31	0.25	0.23	0.33	1.00	
Grit: Perseverance	0.18	0.20	0.15	0.17	0.05	0.22	1.00
Effort in Class	0.29	0.29	0.24	0.24	0.17	0.36	0.57

Notes: n=5610. All correlations are statistically significant at the p<.01 level, except for the correlation between Growth Mindset and Grit: Perseverance, which is statistically significant at the p<.05 level.

This adjustment is implemented by multiplying an estimated correlation between two random variables, *x* and *y*, by the inverse of the square root of the product of the reliability of each measure as follows:

$$r_{xy}^{*} = \frac{\hat{r}_{xy}}{\sqrt{r_{xx} r_{xy}}}$$

I calculate the reliably of the state test score measures by taking the average of the reported test-retest reliabilities in technical manuals for each state across 4th and 5th grade and then averaging these across districts. I estimate Cronbach's alpha reliabilities for the BAM and SAT9-OE as well as for the four social-emotional measures using data from all 4th and 5th grade students who participated in the MET project in Year 2. I report these reliabilities in Table AC2 below.

Table AC2 Estimated Reliabilities	
State Math	0.924
State ELA	0.893
BAM Math	0.716
SAT9-OE Reading	0.851
Growth Mindset	0.780
Grit: Consistency	0.661
Grit: Perseverance	0.692
Effort in Class	0.561

Appendix D

Table AD1: Correlations of Teacher Effects on State Tests, Complex Tasks, and Social-Emotional Measures from Models Using Randomly Assigned Teachers

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	0.59***	1.00					
BAM Math	0.57***	0.33***	1.00				
SAT9-OE Reading	0.32***	0.17**	0.46***	1.00			
Growth Mindset	0.19**	0.18**	0.04	0.16*	1.00		
Grit: Consistency	0.11	0.12	0.07	-0.03	0.22***	1.00	
Grit: Perseverence	-0.10	-0.06	0.07	0.20**	-0.03	0.01	1.00
Effort in Class	-0.02	0.07	0.05	0.12	-0.05	0.01	0.64***

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 229. Classroom effects are derived using the model reported in Column 5 of Table 6.

Appendix E

I can disattenuate the estimated correlations for both sampling and measurement error using an approach analogous to the Spearman (1904) adjustment described in Appendix C. I estimate the reliability of teacher effects for each of the eight outcomes as follows:

$$r_{\tau_j \tau_j} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_{\varepsilon_j}^2}$$

Table 6 provides model-based ML estimate of σ_τ^2 for each outcome. I approximate $\sigma_{\varepsilon_j}^2$ as the average of the squared standard errors of post-hoc predicted BLUE teacher effects from ML models ($\overline{SE_{\tau_j}^2}$). This approach can be seen as providing an upward bound estimate of the true correlations. For example, I find that correlations of approximately 0.60 and above are adjusted to be greater than 1, outside the possible range of correlation coefficients. I report estimated reliabilities for each teacher effect in Table AE1 and disattenuated correlations in Table AE2 below.

Table AE1: Estimated Reliabilities of Teacher Effects

State Math	0.562
State ELA	0.564
BAM Math	0.547
SAT9-OE Reading	0.553
Growth Mindset	0.533
Grit: Consistency	0.508
Grit: Perseverance	0.531
Effort in Class	0.542

Table AE2: Disattentuated Correlations among Teacher Effects on State Tests, Complex Tasks and Social-Emotional Measures

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	1.00	1.00					
BAM Math	1.00	0.65	1.00				
SAT9-OE Reading	0.61	0.45	0.78	1.00			
Growth Mindset	0.42	0.35	0.22	0.41	1.00		
Grit: Consistency	0.34	0.37	0.19	-0.04	0.42	1.00	
Grit: Perseverence	-0.11	-0.04	0.19	0.33	-0.04	0.06	1.00
Effort in Class	0.13	0.14	0.26	0.16	-0.09	0.11	1.00

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. Teacher effects are derived using the model reported in Column 3 of Table 6. Disattentuated estimates outside the range of correlation coefficients are set to 1.

Appendix F

Table AF1: Model-based Restricted Maximium Likelihood Estimates of Teacher Effects on State Tests, Complex Tasks and Social-Emotional Measures without Prior State Test Scores

		Actual Teacher			Randomly Assigned	
	n	(1)	(2)	(3)	(4)	(5)
State Math	4,075	0.178***	0.179***	0.152***	0.157***	0.127***
State ELA	4,074	0.185***	0.189***	0.185***	0.167***	0.162***
BAM Math	3,746	0.173***	0.187***	0.170***	0.166***	0.140***
SAT9-OE Reading	3,766	0.178***	0.186***	0.185***	0.182***	0.177***
Growth Mindset	3,551	0.201***	0.153**	0.141*	0.159**	0.161**
Grit: Consistency	3,473	0.106	0.099	0.105	0.103	0.117*
Grit: Perseverance	3,473	0.155***	0.155**	0.142*	0.154**	0.141*
Effort in Class	3,435	0.169***	0.161**	0.183***	0.119*	.142*
Survey-based Controls			Yes	Yes	Yes	Yes
Peer-level Controls		Yes		Yes		Yes
School FE		Yes				
Randomization Block FE			Yes	Yes	Yes	Yes

Notes: * p<0.05, ** p<0.01, *** p<0.001. Cells report the standard deviation of teacher effect estimates from separate regressions. See notes from Table 6 for further model details.

Table AF2: Correlations of Teacher Effects on State Tests, Complex Tasks, and Socio-Emotional Measures from Models without Prior State Test Scores

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	0.73***	1.00					
BAM Math	0.73***	0.55***	1.00				
SAT9-OE Reading	0.47***	0.40***	0.55***	1.00			
Growth Mindset	0.27***	0.25***	0.20**	0.26***	1.00		
Grit: Consistency	0.37***	0.40***	0.29***	0.14*	0.27***	1.00	
Grit: Perseverence	0.07	0.13	0.20**	0.26***	0.02	0.13	1.00
Effort in Class	0.17*	0.21**	0.23**	0.19**	-0.02	0.15*	0.63***

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. Classroom effects are derived using the model reported in Column 3 of Table 6.