Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies

Matthew A. Kraft
Brown University

February 2016

Updated: July 2017

Forthcoming, *Journal of Human Resources*

Abstract

I exploit the random assignment of class rosters in the MET Project to estimate teacher effects on students' performance on complex 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 outcomes, but weak relationships between these effects and performance measures used in current teacher evaluation systems including value-added to state standardized tests. These findings suggest teacher effectiveness is multidimensional, and high-stakes evaluation decisions are only weakly informed by the degree to which teachers are developing students' complex cognitive skills and social-emotional competencies.

JEL No. H0, I2, J24

Suggested Citation:

Kraft, M.A. (in press). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources*.

Correspondence regarding the paper can be sent to Matthew Kraft at mkraft@brown.edu. PO Box 1983, Brown University, Providence RI, 02912. An early version of this paper was circulated under the title "Teaching for Tomorrow's Economy: Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies." This research was generously supported by the William T. Grant Foundation and the Brown University Undergraduate Teaching and Research Award program. I thank the editor and three anonymous reviewers for their helpful comments as well as seminar participants at Brown, Harvard, Stanford, the University of Connecticut and the Federal Reserve Banks of Boston and New York. I am grateful to Sarah Grace for providing exceptional and extensive research assistance on an early version of the paper as well as to Bruna Lee, Dylan Hogan, and Harry Neuert. All views and errors are my own.

1. Introduction

It is well established that teachers have large effects on students' achievement on state standardized tests (Rockoff 2004; Hanushek and Rivkin 2010; Chetty, Friedman and Rockoff 2014a). However, state tests have typically been narrow measures of student learning, assessing basic literacy and numeracy skills using multiple-choice questions. A review of standardized tests used in 17 states judged as having the most rigorous state assessments found that 98 percent of items on math tests and 78 percent of items on reading tests only required students to recall information and demonstrate basic skills and concepts (Yuan and Le 2012). 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). Chamberlain (2013) found that only one-fifth of teachers' effects on college going were explained by their impacts on standardized tests. Similarly, Jackson (forthcoming) found that teachers' effects on test scores accounted for less than one-third of their effects on high school completion and indicators of college matriculation.

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.[1] Existing research linking teacher effects to outcomes other than traditional standardized assessments has examined three general outcome types: observable behavioral and schooling outcomes such as absences, suspensions, grades, grade retention, and high-school graduation (Jackson forthcoming,

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

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.

The MET Project data allow me to make several important contributions to this literature. First, I estimate teacher effects on a much broader set of student skills and competencies than has been previously examined. In addition to collecting student performance on state standardized tests, MET researchers administered two supplemental achievement tests comprised of open-ended tasks designed to be more direct measures of students' critical thinking and problem-solving skills. 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 educators in recent years.[2] The survey also included a class-specific measure of effort which allows for a direct comparison between teacher effects on global and domain-specific measures of perseverance. I present the first estimates of teacher effects on students' grit, growth mindset, and effort in class. I also provide the first direct evidence of the relationship between teacher effects on state tests, complex open-ended assessments, and social-emotional competencies.

[2] Paul Tough's best-selling book *How Children Succeed* helped to propel 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., 2013a) and the Department of Education has commissioned a paper on "Promoting Grit, Tenacity, and Perseverance" (Shechtman, 2013).

A second key advantage of using the MET data to address these questions is that a subset of teachers participated in an experiment where researchers randomly assigned class rosters among sets of volunteer teachers in the same grades and schools. This design provides the opportunity to identify teacher effects without the strong conditional independence assumption required when using observational data. The extent to which covariate adjustment adequately accounts for nonrandom student sorting when estimating teacher effects on test scores is still a topic of ongoing debate.[3] Even less is known about the validity of this approach for estimating teacher effects on outcomes other than standardized state tests.

Third, the MET data allow me to examine the relationships among teacher effects on an expanded set of student outcomes as well as the primary performance measures used in most teacher evaluation systems. In recent years, states have implemented sweeping reforms to teacher evaluation by adopting more rigorous systems based on multiple measures of teacher effectiveness (Steinberg and Donaldson 2016). I provide among the first evidence on whether the measures used in these high-stakes evaluation systems including value-added to state tests, classroom observations, student surveys, and principal ratings reflect teacher effects on complex cognitive skills and social-emotional competencies.

Leveraging the classroom roster randomization, 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 and 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-

[3] For an overview of the teacher value-added literature see Koedel, Mihaly and Rockoff (2015). For an extensive discussion on the validity of teacher value-added models see Rothstein (2010), Chetty et al. (2014a), and Rothstein's (2017) response to Chetty and his colleagues.

ended tests. While teachers who add the most value to students' performance on state tests in math also appear to strengthen their analytic and problem-solving skills (r=.57), teacher effects on state ELA tests are only moderately correlated with teacher effects on open-ended response items in reading (r=.24). Successfully teaching more basic reading comprehension skills does not indicate that teachers are also developing students' ability to interpret and respond to texts.

Teacher effects on students' social-emotional competencies differ in magnitude, with the largest effects on class-specific effort, the global perseverance subscale of grit, and growth mindset. Comparing the effects of individual teachers across outcomes reveals that correlations between teacher effects on standardized tests and those on social-emotional competencies are never larger than 0.21. Consequently, more than one out of every four teachers who is in the top 25 percent of state test value-added is in the bottom 25 percent of social-emotional value-added. Together, these findings suggest that teacher effectiveness is multiple dimensional and that individual teachers' abilities differ across skillsets.

Turning to teacher evaluation policies, I also find little evidence that performance measures commonly incorporated into high-stakes teacher evaluation systems capture teacher effects on complex cognitive skills or social-emotional competencies. Neither value-added to state standardized tests, scores on classroom observation rubrics, student survey assessments, nor principals' overall assessments of professional practice serve as proxy measures for teacher effects on this broader set of outcomes, either individually or jointly. Correlations between a composite of these teacher performance measures (using commonly applied weights) and teacher effects on social-emotional skills are weak, between .03 and .19. I conclude by discussing the implications of these 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, and 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 math and reading rarely include items designed to assess these abilities (Yuan and Le 2012). Among state tests that do include open-ended ELA questions, these items are often substantially more cognitively demanding tasks than multiple choice questions. However, open-ended items on state math tests typically require students to move beyond recall but rarely require students to solve 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 an emphasis on problem-solving performed better on open-ended math and science tests designed to assess students' decision-making abilities, problem-solving

skills, and conceptual understanding (Le et al. 2006). Using a matched-pair design, researchers at the 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 that 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

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.[4] 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. There exists substantial variation in the racial composition of students across districts such that African-American, Hispanic, and white students each comprise the largest racial/ethnic group in at least one district.

3.1 The Classroom Roster Randomization Experiment

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 percent 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 & 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.[5] 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

[4] Detailed descriptions of the MET data are available at www.metproject.org.
[5] 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).

effect of general education elementary classrooms to minimize potential confounding when students are taught by multiple teachers and outcomes are not class-specific.

3.2 Limitations of the MET Data

While the MET Project has several advantages, the data also have some important limitations. 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 percent of students on these rosters were not taught by 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.

The single year of experimental data combined with my focus on general education elementary classrooms also limits my ability to isolate teacher effects from peer effects and transitory shocks (Chetty et al., 2011). Blazar and 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 percent. I present estimates both with and without peer-level controls to

provide approximate bounds for teacher effects. 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 (forthcoming). 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 I review the causal evidence on interventions targeting these competencies.

3.3 Sample

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, 3) had valid lagged achievement data on state standardized tests in both math and ELA, and 4) were taught by a teacher who is linked with at least five students. These restrictions result in an analytic sample of 4,092 students and 236 general education teachers. Further restricting the analytic sample to require that 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 percent are African-American, 29 percent are Hispanic, 24 percent are white, and 8 percent are Asian. Over 60 percent 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 the teachers hold a graduate degree.

3.4 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. State math and ELA tests for the 4th and 5th grades administered in the six districts in 2011 primarily consisted of multiple-choice items. State test technical manuals suggest that the vast majority of items on these exams assessed students' content knowledge, fundamental reading comprehension, and basic problem-solving skills.[6] Reported reliabilities for these 4th and 5th grade tests in 2011 ranged between 0.85-0.95. In order to make scaled scores comparable across districts, the MET Project converted these scores into rank-based Z-scores.

3.5 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

[6] 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).

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 is comprised 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. The SAT9-OE was developed by Pearson Education and consists of nine open-ended questions about one extended reading passage that tests 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. Similar to state standardized tests, the MET Project converted raw scores on the BAM and SAT9-OE into rank-based Z-scores.

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 Norman Webb's Depth-of-Knowledge framework found that 100 percent of writing, 52 percent of reading, and 24 percent of math open-response items assessed strategic or extended thinking compared to only 32 percent of reading and 0 percent of math multiple-choice items (Yuan and Lee 2014). Demand and wages for jobs that require these complex cognitive skills to perform non-routine tasks, often in combination with strong interpersonal skills, have grown steadily in recent decades (Autor, Levy, and Murnane 2003; Deming 2015; Weinberger 2014).

3.6 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 and 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 0.22 and a disattenuated correlation of 0.33 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, Trzesniewski, and Dweck (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 GPA 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 et al. 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 10th grade students' self-reported answers to survey items that map closely onto the perseverance of effort subscale of grit and 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 the 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, gender, level of parental education, and household income. Grit and growth mindset are generally weaker predictors of outcomes in adulthood compared to measures of academic achievement, but do contain information that is independent from academic ability. 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 disattenuated 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 do not establish an underlying causal relationship or confirm that 4th and 5th graders' self-reported grit and growth mindset have the same predictive power. However, we

do know that grit and growth mindset are negatively correlated with absences and suspensions and positively correlated with GPA among upper elementary and middle school students (West 2016, West et al. 2016). A growing number of randomized control trials evaluating the effect of growth mindset interventions across various grade levels have documented causal effects on short to medium-term academic and behavioral outcomes (Yeager et al. 2014; Miu and Yeager 2015; Paunesku et al. 2015; Yeager et al. 2016). These studies demonstrate that growth mindset 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 et al. 2007) as well as decreased self-reported depressive symptoms (Miu and Yeager 2015) and aggressive desires and hostile intent attributions (Yeager et al. 2013b).

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 goals (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 improved performance and persistence on objective tasks and grades (Alan et al. 2016). Together, these studies suggest that grit and growth mindset are both malleable and likely causal determinants of important intermediary student outcomes for success in later life.

3.7 Achievement Tests, Performance on Complex Tasks, and Social-Emotional Competencies

In Table 3, I present Pearson correlations across the eight outcome measures along with correlations disattenuated for measurement error (see Appendix C for technical details). The clustered patterns of covariance evident in this table illustrate the lack of independence of each

of these measures. Instead, these outcomes likely capture a more limited set of latent constructs. The strongest relationships among the disattenuated correlations 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 measures of 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 0.21 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.

3.8 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 value-added model specifications (Todd and Wolpin 2003; Kane and Staiger 2008; Koedel and Betts 2011; Guarino, Reckase, and Wooldridge 2015; Chetty et al. 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, $\bar{\tau}$, where the weights are determined by the reliability of each estimate.[7] However, variance estimates using fixed effects are biased upward because they conflate true variation with variation due to estimation error. Variance estimates using EB teacher effects 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^2_τ, is bounded between the fixed-effect and EB estimators (Raudenbush and Bryk 2002).

$$\hat{\sigma}^2_{\tau_{EB}} < \sigma^2_\tau < \hat{\sigma}^2_{\tau_{FE}} \quad (1)$$

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 approach 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

[7]Formally, $E\left[\tau_j \middle| \hat{\tau}_j\right] = (1-\lambda_j)\bar{\tau} + (\lambda_j)\hat{\tau}_j$ where $\lambda_j = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_{\varepsilon_j}}$. Here λ_j is the ratio of true teacher variation to total observed teacher variance.

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:

$$Y_{ij} = \alpha_{dg}(f(A_{i,t-1})) + \delta X_i + \beta \bar{A}_{j,t-1} + \theta \bar{X}_j + \pi_{sg} + \varepsilon_{ij} \quad (2)$$

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

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 all polynomial terms with district-by-grade fixed effects. I also include a vector of controls for observable student characteristics (X_i). Student characteristics include indicators for a student's gender, age, race, FRPL status, English language proficiency status, special education status, and participation in a gifted and talented program.[8]

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.[9] Both theory and prior empirical evidence have shown that grades reflect students' cognitive skills as well as social-emotional competencies

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

such as grit and effort (Bowen, Chingos, and 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.

In additional models, I attempt to remove peer effects by controlling for a rich set of average classroom covariates.[10] 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 student's class ($\bar{X}_j$). I present models both with and without peer effects to provide informal upper and lower bounds on the true magnitude of teacher effects. Estimates of the magnitude of teacher effects in a single cross-section where teachers are observed with only one class are likely to be biased upward when peer-level controls are omitted and biased downward when they are included (Kane et al. 2013; Thompson, Guarino, and Wooldridge 2015).[11]

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 restricted maximum likelihood estimation. I

[10] I 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.

[11] In this context where teacher and classroom peer effects are collinear, models that omit peer effects will conflate variation in teacher effect estimates with variation in peer effects across classrooms. The direction and magnitude of bias depends on the correlation between teachers and peer effects. Given the random assignment of class-rosters in the MET data, we would expect estimates of the standard deviation of teacher effects from ML models without peer controls to be inflated. By this same logic, we would expect estimates of teacher effects from ML models with peer controls to over attribute variation in outcomes across classroom to observed peer characteristics. This is because the ML models solve for the coefficients associated with the structural model which include peer measures as the only classroom-level covariates and partition the remaining variance to estimate the magnitude of teacher effects. In application, the direction of bias is not always uniform given noncompliance and the non-random assignment of new students not included in the roster randomization process.

specify τ_j in two different ways – as students' actual teachers and as 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 percent 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).

4. Findings

4.1 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. To further examine this threat, I replicate my primary analyses in samples that exclude gifted and ELL students and report the results in Appendix D. Results are consistent with those reported below with even slightly larger magnitudes of teacher effects.

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.[12] Among the 48 different relationships I test, I find that only one is statistically significant at the 5 percent level. This is consistent with random sampling variation given the number of relationships I test.[13] 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).

4.2 Teacher Effects – Maximum Likelihood Estimates

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

[12] I use value-added estimates calculated by the MET Project 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).

[13] 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).

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 0.18 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 0.17 sd, respectively. Finally, I find suggestive evidence of teacher effects on social-emotional measures ranging from 0.08 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. Comparing results across models with and without peer effects (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 Figure 1 which presents estimates from models with students' actual teachers that condition on peer controls, I find statistically significant effects of broadly similar magnitude (0.14-0.18 sd) across all outcomes except for consistency of interest which is both smaller in magnitude and not statistically significant.

Results from models using students' randomly assigned teachers are slightly attenuated given non-compliance but remain consistent with estimates reported above. Estimates of teacher effects on academic outcomes from models that include peer controls (Column 5) range from 0.11 sd on the BAM to 0.17 sd for the SAT9-OE. Teacher effects on consistency of interest do not achieve statistical significance, while effects on students' growth mindset (0.15 sd),

perseverance (0.14) and effort in class (0.15) are of similar and even slightly larger magnitude than effects on achievement. Together, these results present strong evidence of meaningful teacher effects on students' ability to perform complex tasks and social-emotional competencies.

4.3 Comparing Teacher Effects across Outcomes

I investigate the nature of teacher skills by examining the relationships between individual teachers' effects across the eight outcomes of interest. In Table 7, I present Pearson correlations of the Best Linear Unbiased Estimators (BLUE) of teacher random effects from the ML model that uses students' actual teachers and includes peer controls (Column 3 of Table 6).[14] Correlations among teacher effects from models using randomly assigned teachers produce a consistent pattern of results but are somewhat attenuated due to non-compliance. I present these results in Appendix Table E1.

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 0.58 (Corcoran, Jennings and Beveridge 2012; Goldhaber, Cohen and Walch, 2013; Loeb, Kalogrides and Beteille 2012). Elementary teacher effects on state math tests are also strongly related to their effects on the BAM (0.57). Elementary teachers who are effective at teaching more basic computation and numeracy skills appear to be developing their students' ability to perform complex open-ended tasks in math. This relationship is similar to prior estimates of the correlation between teacher effects on two math exams with more similar content coverage, formats, and levels of cognitive demand (0.64 in Blazar and Kraft, 2017; 0.56 to 0.62 in Corcoran et al., 2012).

[14] These are analogues to empirical Bayes estimates.

In contrast, teacher effects on state ELA exams are a poor proxy for teacher effects on more cognitively demanding open-ended ELA tests. Teacher effects on their students' performance on state standardized exams assessing reading comprehension with multiple-choice items explain less than 6 percent of the variation in teacher effects on the SAT9-OE, an assessment designed to capture students' ability to reason about and respond to an extended passage. The correlation, 0.24, is also notably weaker than prior estimates of the correlation between teacher effects on two different reading exams. Papay (2011) found correlations ranging between 0.44 to 0.58 between a state test in reading and the Scholastic Reading Inventory (SRI).[15] Corcoran and colleagues (2012) found nearly identical correlations (0.44 to 0.58) between teacher effects on the Texas state tests and the Stanford Achievement Test (SAT) in reading.[16] In fact, teachers' value-added to student achievement on the more cognitively demanding open-ended SAT9-OE reading exam is most strongly related to their effects on the similarly demanding open-ended BAM math test (0.46) than with their value-added to state ELA tests.

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 open-ended 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 0.10 and 0.21. Teachers' ability to motivate their students' perseverance and effort is consistently a stronger predictor of teacher effects on students'

[15] Papay (2011) finds much lower correlations between the state test and the (SAT) in reading (.15 to .36) and the SRI and SAT in reading (.23 to .40). However, the SAT was administered in the fall likely confounding teacher effect estimates in time *t* with both differential summer learning and, to a lesser degree, a student's teacher in time *t*+1. The correlations I report in the text are based on exams that were both given in the spring.
[16] Corcoran et al. (2012) report that the state exams and the SAT in reading were administered "at roughly the same time of year" (p.4).

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, with a correlation of 0.61. This 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 2. This relationship captures the strongest correlation I observe between teacher effects on social-emotional competencies and state tests (0.21). A total of 42 percent of teachers in the sample have above average effects on one outcome but below average effects on the other (21 percent in quadrant II and 21 percent in quadrant IV). Only 28 percent of teachers have effects that are above average for both state math tests and growth mindset (quadrant I). The proportion of teachers who have above average effects on both state math tests and other social-emotional measures is even lower. These findings illustrate how teachers are not simply "effective" or "ineffective" but instead have abilities that may differ across multiple dimensions of effectiveness.

4.4 Assessing Potential Bias in Teacher Effect Correlations

The pairwise correlations presented in Table 7 are imperfect estimates of the true relationships between teacher effects, although the net direction of potential biases is not obvious. Noise in teacher effect estimates due to the imperfect reliability of student outcome

measures will bias estimates downward.[17] At the same time, class-specific shocks and unobserved student traits correlated with multiple outcomes can induce an upward bias. I explore the magnitude of potential biases by estimating upper and lower bounds for these correlations.

I first estimate upper bounds for Table 7 by disattenutating estimates for measurement error using an approach analogous to the Spearman (1904) correction described in Appendix C. I provide technical details for this procedure and report the results in Appendix G. The low estimated reliabilities of teacher effect estimates (0.48 to 0.59) result in almost a doubling of the magnitude of the unadjusted correlations with some correlations disattenuated to be greater than 1, outside the possible range of correlation coefficients. This is because the Spearman adjustment assumes that errors in both measures are uncorrelated with each other, an assumption likely violated in this setting given that teacher effects are estimated using the same classroom of students across outcomes. Even these extreme upper bound estimates show that correlations between teacher effects on state tests and social-emotional competencies are never larger than 0.42 (state math and growth mindset).

I next estimate lower bounds for Table 7 by examining correlations among teacher effects from different years for a subset of outcomes available in both years.[18] This approach purges correlations of upward bias introduced by correlated errors from a common estimation sample. In all years, I use teacher effect estimates calculated by the MET Project using a standard covariate-adjustment model (Kane and Cantrell 2010) to hold the modeling approach constant. Appendix

[17] I also examine the degree to which sampling error may attenuate these correlation coefficients by estimating the sensitivity of my estimates to class size. I present the results in Appendix F. These findings suggest the post-hoc predicted BLUE random effect estimates I use when correlating teacher effects sufficiently correct for sampling error due to small class sizes.

[18] This approach eliminates the direct correlation between the errors of individual students across outcomes, but is still susceptible to differential sorting patterns among teachers that are stable across classes or years. It also implicitly assumes an individual teacher's effect does not change over time or differ based on class or school characteristics.

Table G3 compares correlation coefficients calculated among the analytic sample of general elementary school teachers based on estimates for the same year (Panel A) and estimates in different years (Panel B). Consistent with prior studies, I find that teacher effect correlations estimated from the same class are inflated upwards, sometimes substantially, relative to teacher effects from across years (Goldhaber et al. 2013; Kane and Cantrell 2010). The largest degree of upward bias 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 biased upward to a slightly lesser degree. Still, the patterns in these lower bound estimates remain the same; correlations between teacher effects on state tests of different subjects are the largest (0.26), followed by correlations between effects on state tests and open-ended tests (0.06-0.17), and finally correlations between social-emotional competencies and test scores (0.0-0.12).

While it is difficult to know how these biases interact, I interpret these findings to suggest that attenuation bias due to noise in teacher effects is largely if not completely offset by the upward bias due to correlated errors caused by a common classroom sample. I expect the results reported in Table 7 may slightly underestimate the true magnitude of these correlations but support general inferences about the relative magnitude of these correlations across outcomes.

4.5 Do Teacher Evaluation Systems Capture 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 systems that

incorporate multiple measures including estimates of contributions to student learning, classroom observation scores, student surveys, and assessments of professional conduct (Steinberg and Donaldson 2016). Teachers' evaluation ratings are typically constructed from a weighted combination of these measures. Classroom observations nearly always account for the largest percentage of the overall score, although the weights assigned to measures vary meaningfully across districts and states (Steinberg and Kraft forthcoming).

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 8, 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. At the same time, the relationships between performance measures and true teacher effects is likely somewhat stronger than the estimates reported in Table 8 which rely on imprecise measures from a single year (Kane and Staiger 2012). I compare my teacher effect estimates with the most common metrics used in teacher evaluation systems: value-added in math and ELA[19]; ratings on two widely used classroom observation instruments, the Classroom Assessment Scoring System (CLASS) and the Framework for Teaching (FFT); students' opinions of their teachers' instruction captured on the TRIPOD survey (Kane and Cantrell 2010); and principals' overall ratings of teachers'

[19] This value-added performance measure estimate differs from my teacher effect estimates in several ways. It is the average of teacher effect estimates in math and reading calculated by the MET Project using a standard covariate adjustment model and including all students in teachers' classes regardless of whether students were part of the roster-randomization study (see Bill & Melinda Gates Foundation 2013). Similar to Table 7, teacher effect estimates are post-hoc predicted BLUE random effect estimates derived from a model using students' actual teachers and controlling for classroom peer characteristics. The estimation sample is limited to students who were included in the roster randomization process as described in section 3.3.

performance using a six-point Likert scale ranging from "Very Poor" to "Excellent."

I find that neither value-added scores, classroom observation scores, student surveys, nor principal ratings serve as close proxies for teacher effects on complex cognitive skills or social-emotional competencies. Principal ratings have the strongest relationship with teacher effects on growth mindset with a correlation of 0.17. In aggregate, classroom observations scores do not appear to reflect teacher effects on this broader set of outcomes despite the wide range of domains covered by these rubrics. In supplemental analyses, I find that the strongest correlation across all eight teacher effects and the 12 CLASS domains is .16 (p=.02) between teacher effects on effort in class and the "Productivity" domain. The strongest correlation with the eight FFT domains is .17 (p=.01) between teacher effects on growth mindset and the "Establishing a Culture for Learning" domain. 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.

I illustrate how summative teacher ratings from high-stakes teacher evaluation systems compare to the teacher effects I estimate by constructing proxy summative scores for teachers using the performance measures described above. I calculate scores using a weighted linear sum of value-added, observation, student, and principal ratings, with weights that reflect a prototypical evaluation system for teachers in tested grades and subjects.[20] As show in Table 8, teachers' summative ratings are only weakly related their ability to develop students' complex

[20] I draw upon evidence from Steinberg and Donaldson (2016) to select metrics and weights. I standardize all four performance measures to be mean zero and have a variance of one and then add them using the following weights: $Score = .50 * CLASS + .35 * ValueAdded + .05 * Survey + .10 * Principal\ Rating$. Results using FFT in place of CLASS as well as alternative weights produce similar results.

cognitive skills and social-emotional competencies. The two strongest relationships are with teacher effects on open-ended tasks in math and growth mindset, with correlations of .19. Among teachers ranked in the bottom fourth of the evaluation ratings, I estimate that 27 percent are actually in the top quartile of teacher effects on complex math tasks and 21 percent are in the top quartile of effects on growth mindset. 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 Falsification Tests & Differential Reliability Across Measures

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 conduct several falsification tests for spurious results and find no compelling evidence that the results are driven by sampling error. First, I generate 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 repeat this process 100 times. I report the average of these simulated results in Panel A of Table 9. The estimates across models are quite small, between 0.03 and 0.04 standard deviations.

I next test for teacher effects on a range of student characteristics that should be unaffected by teachers. 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 9 Panel B, I easily reject teacher effects across all of these measures except age for models using students' actual teachers.

In Table 9 Panel C, I further demonstrate that ML estimates are not driven by unexplained variance due to the lower reliability of open-ended tests or survey scales. I test this by, ex post, randomly reassigning students to teachers in the analytic sample in a way that 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. Averaging estimates across 100 repeated random draws, I find that the majority of estimates converge to precise zeros. Only estimates for consistency are of meaningful magnitude (0.08), but this is of less concern given that I fail to find any significant effects on this outcome across the primary analyses. Together, these falsification tests lend strong support to the validity of the teacher effect estimates.

5.2 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).[21] In this context, the MET Project's experimental design restricts the identifying variation to within school-grade cells, limiting the potential for reference bias at the school-level and grade-level within a school. Additional empirical tests provide further evidence against reference bias as a primary driver of the main results. Following West et al. (2016), I examine how the direction and

[21] 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 2015).

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.[22]

As shown in Table 10, 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.

5.3 Removing Prior Test Scores

Across all models, I include prior achievement scores from state tests along with additional controls for student (and peer) characteristics that 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 scores but not prior scores on open-ended tests or social-emotional competencies creates an asymmetry in that only models with state test scores as outcomes include corresponding controls for prior outcome measures. However, unlike prior

[22] 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 regions (Schmitt et al. 2007). Notably, they 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.

approaches which rely primarily on lagged test scores, my identification strategy leverages the random assignment of class rosters to address student sorting. I examine the sensitivity of the ML variance estimates from Table 6 and correlations across teacher effects from Table 7 by comparing them to estimates from models that exclude controls for prior test scores as well as peer average test scores.

Teacher effect estimates that omit prior scores presented in Appendix H are slightly larger, likely due to the between-classroom variance in a randomization block that was previously accounted for by conditioning on individual and peer-average prior achievement. Results from models that include peer controls increase the most, between 0 to 35 percent, 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. The inflated magnitude of these correlations is likely due to an increase in correlated errors among teacher effects which prior test scores helped to reduce. Overall, these results suggest the primary findings are not driven by the asymmetric set of lagged outcome measures.

5.4 Teacher Effects – Upper and Lower Bound Average Residual Estimates

As a robustness check for my preferred model-based ML estimation approach, I also estimate upper and lower bounds for the variance of teacher effects using a two-step estimation approach following Kane et al. (2013). This allows me to relax the random effects normality assumption necessary for equation (2). 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 (2), omitting teacher random effects. In a second step, I average student residuals at the teacher level to estimate teacher effects. The variance of these average classroom

residuals produces the upper bound estimates reported in Panel A of Appendix Table I1. I then shrink the average classroom residuals as described in footnote 7.[23] The variance of these shrunken EB teacher effects provide lower-bound estimates reported in Panel B of Table I1.

Estimated bounds conform to the ex-ante predictions described in section 3.8 and almost uniformly contain my preferred estimates in Table 6. As expected, unshrunken average residuals overstate the effects of teachers while shrunken average residuals understate the magnitude of these effects. Unshrunken teacher effects on open-ended tasks and social-emotional measures are all larger than those on state tests whereas before they were of similar magnitude. Unlike the ML estimates, average residuals are biased differentially because outcomes with lower reliability and more measurement error have more unexplained variance across classrooms. Those measures with the highest reliabilities are closest to the preferred ML estimates. Shrunken average residual estimates produce lower bounds that in some cases converge to zero. These estimates are quite conservative given the small student sample sizes for a single elementary school classroom result in low reliabilities for individual estimates which are then shrunken substantially towards the grand mean of zero. Overall, these results confirm that our qualitative findings are not a product of the identifying assumptions of the model-based ML estimation process.

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

[23] 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 $\sigma^2_{\bar{\varepsilon}_j}$ is the squared standard error of teacher j's average classroom residuals ($SE^2_{\bar{\varepsilon}_{ij}}$). I calculate standard errors using 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.

on state standardized tests. 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 full set of student abilities and competencies that predict positive adult outcomes. Questions remain about whether teachers and schools that are judged as effective by state standardized tests are also developing students' more complex cognitive skills and social-emotional competencies. This study suggests that this is often not the case.

The large differences in teachers' ability to raise student performance on achievement tests (Chetty et al. 2014a; Hanushek and Rivkin 2010) and the inequitable distribution of those teachers who are most successful at raising achievement (Clotfelter, Ladd, Vigdor 2006; Goldhaber, Lavery, and Theobald 2015; 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' contributions to students' performance on state tests in math are strong proxies for their effects on students' abilities 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 they are developing students' social-emotional competencies.

Teaching core academic skills along with social-emotional competencies and the ability to perform unstructured tasks need not be competing priorities in a zero-sum game. I find that the relationships between teacher effects across this expanded set of student outcomes are

consistently positive although often weak. As these analyses demonstrate, there are teachers who teach core academic subjects in ways that also develop students' complex problem-solving skills and social-emotional competencies. I find that about 8 percent of teachers are rated in the top 25 percent of both value-added to complex cognitive skills and social-emotional competencies. Roughly 3 percent of teachers are in the top quartile of value added to state tests, complex cognitive skills and social-emotional competencies. Going forward, we need to know more about the types of curriculum, instruction, organizational practices, and school climates that allow 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 value-added to state states, observation scores, student surveys, nor principal ratings serve as close proxies for teacher effects on important skills and competencies not captured by state tests. Between one out of every four to six teachers who are rated among the top 10% based on a weighted composite of commonly used performance measures has below average effects on complex problem-solving skills and social-emotional competencies. In recent years, dozens of states have adopted new assessments aligned with the Common Core State Standards that move in the direction of assessing more complex cognitive skills (Doorey and Polikoff 2016). While these assessments may better align incentives for teachers, they face several challenges including the traditionally lower reliability and higher cost of scoring constructed response items, increasing political opposition, and public pushback to higher standards that result in fewer students scoring at proficient or advanced levels. The long-term success of these reforms may ultimately be

determined by the degree to which teachers receive the support they need to adapt their teaching to help students meet the demands of these higher standards.

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. While there exists potential to improve the reliability and robustness of these measures, it may be that observable student outcomes such as GPA, grade retention, attendance and disciplinary incidents are ultimately more tractable measures for policy purposes (Whitehurst 2016). Persistent measurement challenges and the susceptibility of even these observable measures to manipulation may mean that it is more productive to focus on formative assessment approaches that help promote a dialogue among teachers, parents and students about the importance of social-emotional development. As Albert 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

Alan, Sule, Teodora Boneva, and Seda Ertac. 2016. “Ever failed, try again, succeed better: Results from a randomized educational intervention on grit.” HCEO Working Paper.

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim D. Kautz. 2011. “Personality Psychology and Economics.” *Handbook of the Economics of Education*. 4:1-181.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics. 118(4):1279-1334.*

Bill & Melinda Gates Foundation. 2013. User Guide to Measures of Effective Teaching Longitudinal Database (MET LDB). Inter-University Consortium for Political and Social Research.

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

Blazar, David., Matthew A. Kraft. 2017. “Teacher and Teaching Effects on Students’ Attitudes and Behaviors.” *Educational Evaluation and Policy Analysis* 39*:*146-170.

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

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

Bowles, Samuel, Herbert Gintis, Melissa Osborne. 2011. “The Determinants of Earnings: A Behavioral Approach.” *Journal of Economic Literature* 39:137-176.

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

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

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

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

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

Deming, David J. 2015. "The Growing Importance of Social Skills in the Labor Market." NBER Working Paper 21473.

Dobbie, Will, and Roland G. Fryer Jr. 2015. "The Medium-Term Impacts of High-Achieving Charter Schools on Non-Test Score Outcomes." Journal of Political Economy, 123(5):985-1037

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

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

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

Duckworth, Angela Lee, Heidi Grant, Benamin Loew, Gabriele Oettingen, and Peter M. Gollwitzer. 2011."Self- regulation Strategies Improve Self- discipline in Adolescents: Benefits of Mental Contrasting and Implementation Intentions." *Educational Psychology* 31: 17-26.

Duckworth, Angela Lee, Teri A. Kirby, Anton Gollwitzer and Gabriele Oettingen.2013."From Fantasy to Action: Mental Contrasting with Implementation Intentions (MCII) Improves Academic Performance in Children." *Social Psychological and Personality Science* 4:745-753.

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

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

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

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

Goldhaber, Dan, James Cowan, and Joe Walch. 2013. "Is a good elementary teacher always good? Assessing teacher performance estimates across subjects." *Economics of Education Review* 36:216-228.

Goldhaber, Dan, Lesley Lavery, and Roddy Theobald. 2015."Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students." *Educational Researcher* 44(5):293-307.

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

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

Heckman, James J., Rodrigo Pinto, and Peter A. Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103(6): 2052–2086

Jackson, C. Kirabo. forthcoming. "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes." Journal of Political Economy.

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

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

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

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

Kane, Thomas J., Daniel F. McCaffrey, Tre Miller, and Douglas O. Staiger. 2013. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." *Bill & Melinda Gates Foundation.*

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

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

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

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015 "Value-added modeling: A review." *Economics of Education Review* 47: 180-195.

Ladd, Helen F. and Lucy C. Sorensen. 2017. "Returns to Teacher Experience: Student Achievement and Motivation in Middle School." *Education Finance and Policy* p. 241-279.

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

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

Loeb, Susanna, Demetra Kalogrides, and Tara Béteille. 2012. "Effective schools: Teacher hiring, assignment, development, and retention." *Education Finance and Policy* 7:269-304.

Miu, Adriana Sum and David Scott Yeager. 2015. "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*: 726-743.

MET Project. 2009. "Memorandum: MET Test Recommendations."

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

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

OECD. 2013. PISA 2015: Draft Collaborative Problem Solving Framework.

Papay, John P. 2011. "Different tests, different answers: The stability of teacher value-added estimates across outcome measures." *American Educational Research Journal* 48: 163-193.

Paunesku, David, Gregory M. Walton, Carissa Romero, Eric N. Smith, David S. Yeager, and Carol S. Dweck. 2015. "Mind-set Interventions are a Scalable Treatment for Academic Underachievement." *Psychological Science* 26: 784–793.

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

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

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

Rothstein, Jesse. 2017. "Measuring the Impact of Teachers: Comment." *American Economic Review* 107: 1656-1684.

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

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

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

Spearman, Charles. 1904. “The Proof and Measurement of Association between Two Things,” *The American Journal of Psychology* 15: 72-101.

Steinberg, Matthew P. and Morgaen Donaldson. 2016. "The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era." *Education Finance and Policy* 11(3): 340-359.

Steinberg, Matthew P., and Matthew A. Kraft. Forthcoming. "The Sensitivity of Teacher Performance Ratings to the Design of Teacher Evaluation Systems." *Educational Researcher*.

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

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

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

Van Merriënboer, Jeroen J. G. 1997. *Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training.* Educational Technology Publications. Englewood Cliffs, NJ.

Weinberger, Catherine J. 2014. “The Increasing Complementarity between Cognitive and Social Skills” *Review of Economics and Statistics* 96: 849-861.

West, Martin R. 2016. "Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California’s CORE districts."

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

Whitehurst, Grover J. 2016. "Hard Thinking on Soft Skills." *Evidence Speaks Reports*. Brookings Institute. 1(14): 5.

Yeager, David Scott, Dave Paunesku, Gregory M. Walton, and Carol S. Dweck. 2013a. “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."*

Yeager, David Scott, Adriana S. Miu, Joseph Powers, and Carol S. Dweck. 2013b. "Implicit Theories of Personality and Attributions of Hostile Intent: A Meta-analysis, an Experiment, and a Longitudinal Intervention." *Child development* 84:1651-1667.

Yeager, David Scott, Rebecca Johnson, Brian James Spitzer, Kali H. Trzesniewski, Joseph Powers, and Carol S. Dweck. 2014.“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*:* 867.

Yeager, David S., Gregory M. Walton, Shannon T. Brady, Ezgi N. Akcinar, David Paunesku, Laura Keane, L. et al. 2016. “Teaching a Lay Theory Before College Narrows Achievement Gaps at Scale”. *Proceedings of the National Academy of Sciences* 113:E3341-E3348.

Yuan, Kun, and Vi-Nhuan Le. 2012. “Estimating the Percentage of Students Who Were Tested on Cognitively Demanding Items through the State Achievement Tests.” *RAND Corporation*. Santa Monica, CA.

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

Zeiser, Kristina L., James Taylor, Jordan Rickles, Michael S. Garet, Michael Segeritz. 2014. *Evidence of Deeper Learning. Findings from the Study of Deeper Learning: Opportunities and Outcomes*. American Institutes for Research,

Figures

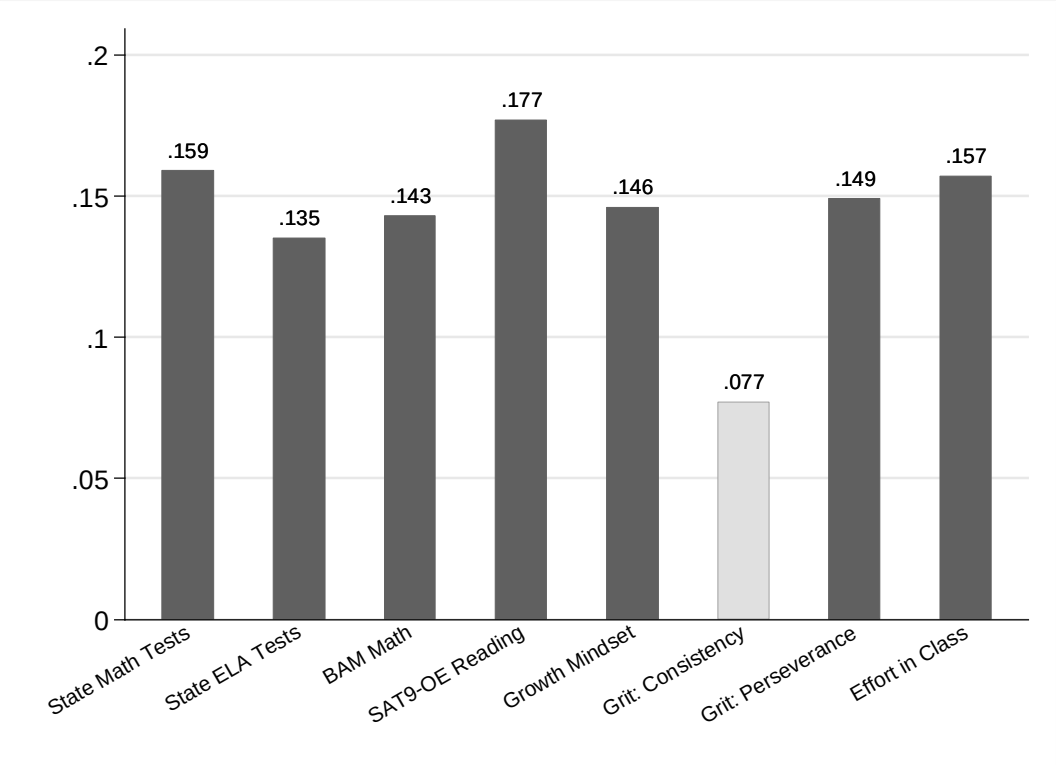


Figure 1: The magnitude of teacher effect estimates on state tests, complex open-ended assessments, and social-emotional competencies. All estimates are statistically significant at the 0.05 level (dark grey bars) except for Grit: Consistency (light grey bar).

Notes: Estimates are model-based restricted maximum likelihood estimates of teacher effects using students' actual teachers and controlling for classroom peer characteristics (Column 3 of Table 6) with samples ranging from 3,435 to 4,075.

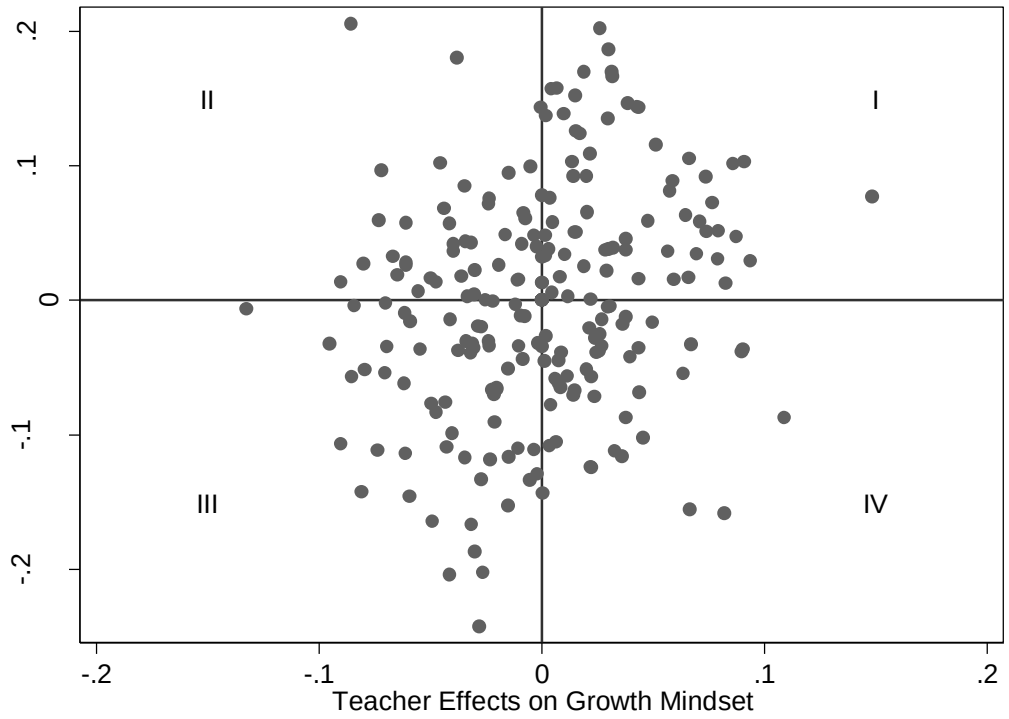


Figure 2: Scatterplot of teacher effects on state math tests and growth mindset from empirical Bayes estimates (n=228). The scatterplot represents a correlation of .21.

Notes: Empirical Bayes estimates are the Best Linear Unbiased Estimators of teacher random effects from ML models that uses students' actual teachers and includes peer controls (Column 3 of Table 6). The scales of both teacher effect estimates are measured in student-level standard deviation unites of the outcomes.

Tables

Table 1: Student and Teacher Characteristics

	Students			Teachers	
	Analytic Sample	U.S. Public Schools in Cities	U.S. Public Schools	Analytic Sample	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 analytic 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 are 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 from the Educational Longitudinal Study.

	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***	3125.5***	-0.027***	0.005	0.070***	0.073***
	(0.006)	(0.007)	(341.1)	(0.004)	(0.007)	(0.007)	(0.007)
Grit: Perseverance of Effort	0.058***	0.026***	1631.6***	-0.008*	0.019**	0.035***	0.036***
	(0.006)	(0.006)	(313.7)	(0.003)	(0.006)	(0.006)	(0.006)
Growth Mindset in Math	0.011*	0.006	848.2**	-0.006*	-0.009	0.019**	0.008
	(0.005)	(0.006)	(324.2)	(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. Cells contain estimates and associated standard errors from OLS models where adult outcomes at age 26 in 2012 are modeled as a function of academic achievement and self-reported measures of grit and growth mindset assessed in 10th grade. Academic achievement is a composite measure of students' academic ability in math and reading based on students' scores on multiple-choice achievement tests in 10th grade. Measures of grit and growth mindset are proxy measures constructed from student survey questions similar to the original scales. 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. See Appendix B for further details about the measures.

Table 3: Student-Level 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
Panel A: Raw Correlations							
State ELA	0.74						
BAM Math	0.66	0.58					
SAT9-OE Reading	0.43	0.49	0.54				
Growth Mindset	0.25	0.29	0.26	0.23			
Grit: Consistency	0.27	0.31	0.25	0.23	0.33		
Grit: Perseverance	0.18	0.20	0.15	0.17	0.05	0.22	
Effort in Class	0.29	0.29	0.24	0.24	0.17	0.36	0.57
Panel B: Disattentuated Correlations							
State ELA	0.81						
BAM Math	0.81	0.73					
SAT9-OE Reading	0.49	0.56	0.69				
Growth Mindset	0.29	0.35	0.35	0.28			
Grit: Consistency	0.35	0.40	0.36	0.31	0.46		
Grit: Perseverance	0.23	0.25	0.21	0.22	0.07	0.33	
Effort in Class	0.40	0.41	0.38	0.35	0.26	0.59	0.91

Notes: n=5610. Panel A reports raw Pearson product-moment correlations from a sample which includes all 4th and 5th grade students of general elementary school teachers with complete outcome data. Panel B reports these same correlations adjusted for attenuation bias due to measurement error following Spearman (1904). All correlations in Panel A 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: Tests for Post-Attrition Randomization Balance in Student Demographic Characteristics 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 calculated by the MET Project using a standard covariate adjustment model and can be interpreted in student standard deviation units. The standard deviation of the teacher effects from prior years are .226 for math, .170 for ELA, .211 for BAM, and .255 for SAT9-OE.

Table 6: Model-based Restricted Maximum 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	0.175***	0.159***	0.150***	0.139***	0.124***
State ELA	4,074	0.142***	0.135***	0.137***	0.123***	0.123***
BAM Math	3,746	0.137***	0.143***	0.126***	0.129***	0.110**
SAT9-OE Reading	3,766	0.168***	0.177***	0.174***	0.176***	0.169***
Growth Mindset	3,551	0.196***	0.146**	0.133*	0.159***	0.154**
Grit: Consistency	3,473	0.082	0.077	0.078	0.080	0.102
Grit: Perseverance	3,473	0.149**	0.149**	0.136*	0.153**	0.138*
Effort in Class	3,435	0.162***	0.157**	0.183***	0.115*	0.149**
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 estimates of the standard deviation of teacher effects from separate regressions. Columns 1 through 3 estimate the effect of 4th and 5th grade students' actual teacher while columns 4 and 5 estimate intent-to-treat effects of the teachers students were randomly assigned to via the MET classroom roster randomization process. All models include controls for students' prior achievement in math and reading, 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-OE Reading	Growth Mindset	Grit: Consistency	Grit: Perseverance
State ELA	0.58***						
BAM Math	0.57***	0.31***					
SAT9-OE Reading	0.38***	0.24***	0.46***				
Growth Mindset	0.21***	0.12	0.10	0.19**			
Grit: Consistency	0.17*	0.21**	0.05	-0.03	0.22***		
Grit: Perseverance	-0.04	0.00	0.10	0.19**	-0.03	0.04	
Effort in Class	0.05	0.09	0.14*	0.10	-0.08	0.06	0.61***

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. This table reports unadjusted Pearson product-moment correlations using post-hoc predicted BLUE teacher random effect estimates derived from a model using students' actual teachers and controlling for classroom peer characteristics (Column 3 of Table 6).

Table 8: Correlations of Teacher Performance Measures with Teacher Effects on State Tests, Complex Tasks, and Social-Emotional Measures

	Composite Evaluation Score	Value-Added in Math & ELA		CLASS		FFT		Student Surveys		Principal Ratings
	Weights	35%	na	50%	na	na	na	5%	na	10%
	Current Year	Current Year	Prior Year	Current Year	Prior Year	Current Year	Prior Year	Current Year	Prior Year	Current Year
State Math	0.29***	0.51***	0.21**	0.07	0.00	0.07	0.00	0.01	0.03	0.16*
State ELA	0.27***	0.49***	0.14*	0.04	0.06	0.11	0.02	0.09	0.14	0.10
BAM Math	0.19**	0.28***	0.13*	0.08	0.02	0.12	0.06	0.14*	0.11	0.08
SAT9-OE Reading	0.18*	0.19**	-0.00	0.10	0.07	0.12	0.02	0.05	0.08	0.05
Growth Mindset	0.19*	0.14*	0.14*	0.12	0.07	0.10	0.07	0.01	0.10	0.17*
Grit: Consistency	0.15*	0.12	0.04	0.04	0.06	0.06	0.03	0.09	0.00	0.05
Grit: Perseverance	0.03	0.00	-0.09	0.06	0.02	0.07	-0.05	0.19**	-0.04	-0.13
Effort in Class	0.10	0.05	-0.04	0.12	0.06	0.13	-0.02	0.20**	0.02	-0.10

Notes: *p<0.05; **p<0.01; ***p<0.001. n=227 except for correlations with composite evaluation scores or principal ratings (n=190). This table reports unadjusted Pearson product-moment correlations. Teacher effects are post-hoc predicted BLUE random effect estimates derived from a model using students' actual teachers and controlling for classroom peer characteristics (Column 3 of Table 6). Value-Added in Math and ELA is the average of teacher effect estimates in math and ELA calculated by the MET Project using a standard covariate adjustment model and including all students in teachers' classes regardless of whether students were part of the roster-randomization study. Scores from the Classroom Assessment Scoring System (CLASS) and Framework for Teaching (FFT) are calculated using the first factor from a Principal Component Analysis of the average domain-level scores across observations from each instrument. Student surveys are teacher-level averages of students' average responses to the TRIPOD seven C's survey items for all students with no missing responses. Principal ratings are from a single survey item asking principals to rate teachers on a six point Likert scale ranging from Very Poor to Exceptional.

Table 9: Falsification Tests of Teacher Effects

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
		Panel A: Random Numbers as Outcome				
Random Number	4,092	0.032	0.031	0.033	0.036	0.036
		Panel B: Student Characteristics that Should be Unaffected by Teachers				
Male	4,092	0.000	0.000	0.000	0.000	0.000
Age	4,092	0.050*	0.048*	0.051*	0.044	0.042
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	0.026	0.019	0.021	0.000	0.000
Hispanic	4,092	0.022	0.031	0.028	0.029	0.024
		Panel C: Outcomes of Students Re-randomized to Teachers				
State Math	4,075	0.000	0.024	0.016	-	-
State ELA	4,074	0.000	0.035	0.015	-	-
BAM Math	3,723	0.000	0.000	0.000	-	-
SAT9-OE Reading	3,753	0.000	0.000	0.000	-	-
Growth Mindset	3,547	0.000	0.000	0.000	-	-
Grit: Consistency	3,463	0.082	0.082	0.081	-	-
Grit: Perseverance	3,463	0.000	0.000	0.000	-	-
Effort in Class	3,435	0.000	0.018	0.000	-	-
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. Panel A reports results of the "effect" of teachers on random numbers drawn from the standard normal distribution, averaged across 100 repeated random draws. Panel B reports estimates of teacher effects on their own or randomly assigned students' demographic characteristics. Panel C reports estimated teacher effects on my primary outcomes of interest when students are randomly assigned to teachers, averaged across 100 repeated random draws. All results are post-hoc predicted BLUE teacher random effect estimates derived from a model using students' actual teachers and controlling for classroom peer characteristics (Column 3 of Table 6) from separate regressions. The sample size for Free or Reduced Price Lunch is limited because one participating district did not provide this information. See notes from Table 6 for further model details.

Table 10: 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	0.05**	0.16*	0.15	0.09***	0.19**	0.37**
Grit: Consistency	0.07***	0.19**	0.33**	0.09***	0.21**	0.15
Grit: Perseverance	0.04**	0.06	0.37**	0.07***	0.08	0.25*
Effort in Class	0.09***	0.23**	0.55***	0.08***	0.23***	0.34**
n	3543	230	67	3543	230	67

Notes: *p<0.05; **p<0.01; ***p<0.001. This table reports unadjusted Pearson product-moment correlations. Student-level correlations are estimated in a student-level dataset. Class- and school-level estimates are estimated using classroom and school averages, respectively. Test scores gains are the residuals from regressions of students' current score on cubic functions of their prior math and ELA state test scores. Reported sample sizes represent the largest sample among the four estimates in each column.

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" assigned values 1 through 4, respectively. For both variables, indices were created by averaging the responses to all sub-questions identified as pertaining to growth mindset and perseverance 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 perseverance, 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 Panel B by disattenuating the raw correlation coefficients in Panel A using the Spearman (1904) adjustment. 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 C1 below.

Table C1 Estimated Reliabilities of Outcome Measures

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 D1: Model-based Restricted Maximum Likelihood Estimates of Teacher Effects on State Tests, Complex Tasks and Social-Emotional Measures Excluding Gifted Students and English Language Learners (ELLs)

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
		Panel A: Results Excluding Gifted Students and ELLs				
State Math	3,117	0.194***	0.164***	0.156***	0.147***	0.137***
State ELA	3,120	0.156***	0.149***	0.148***	0.139***	0.137***
BAM Math	2,847	0.158***	0.161***	0.152***	0.137***	0.131***
SAT9-OE Reading	2,861	0.199***	0.197***	0.202***	0.196***	0.195***
Growth Mindset	2,697	0.232***	0.179**	0.177**	0.180**	0.189*
Grit: Consistency	2,633	0.074	0.099	0.069	0.081	0.059
Grit: Perseverance	2,633	0.191***	0.185***	0.191**	0.172**	0.176**
Effort in Class	2,602	0.195***	0.207***	0.246***	0.165**	0.216***
		Panel B: Results Including Gifted Students and ELLs from Table 6				
State Math	4,075	0.175***	0.159***	0.150***	0.139***	0.124***
State ELA	4,074	0.142***	0.135***	0.137***	0.123***	0.123***
BAM Math	3,746	0.137***	0.143***	0.126***	0.129***	0.110**
SAT9-OE Reading	3,766	0.168***	0.177***	0.174***	0.176***	0.169***
Growth Mindset	3,551	0.196***	0.146**	0.133*	0.159***	0.154**
Grit: Consistency	3,473	0.082	0.077	0.078	0.080	0.102
Grit: Perseverance	3,473	0.149**	0.149**	0.136*	0.153**	0.138*
Effort in Class	3,435	0.162***	0.157**	0.183***	0.115*	0.149**
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 estimates of the standard deviation of teacher effects from separate regressions. Columns 1 through 3 estimate the effect of 4th and 5th grade students' actual teacher while columns 4 and 5 estimate intent-to-treat effects of the teachers students were randomly assigned to via the MET classroom roster randomization process. All models include controls for students' prior achievement in math and reading, 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 D2: Correlations of Teacher Effects on State Tests, Complex Tasks, and Socio-Emotional Measures in a Sample Excluding Gifted Students and English Language Learners (ELLs)

	State Math	State ELA	BAM Math	SAT9-OE Reading	Growth Mindset	Grit: Consistency	Grit: Perseverance
	Panel A: Results Excluding Gifted Students and ELLs						
State ELA	0.55***						
BAM Math	0.61***	0.36***					
SAT9-OE Reading	0.44***	0.29***	0.52***				
Growth Mindset	0.20**	0.06	0.06	0.15*			
Grit: Consistency	0.25***	0.18*	0.11	0.10	0.20**		
Grit: Perseverance	-0.08	-0.02	0.23**	0.28***	0.05	0.08	
Effort in Class	0.02	0.06	0.24**	0.17*	0.00	0.03	0.62***
	Panel B: Results Using Full Analytic Sample from Table 7						
State ELA	0.58***						
BAM Math	0.57***	0.31***					
SAT9-OE Reading	0.38***	0.24***	0.46***				
Growth Mindset	0.21***	0.12	0.10	0.19**			
Grit: Consistency	0.17*	0.21**	0.05	-0.03	0.22***		
Grit: Perseverance	-0.04	0.00	0.10	0.19**	-0.03	0.04	
Effort in Class	0.05	0.09	0.14*	0.10	-0.08	0.06	0.61***

Notes: *p<0.05; **p<0.01; ***p<0.001. n=207 in Panel A and n=227 in Panel B. This table reports unadjusted Pearson product-moment correlations using post-hoc predicted BLUE teacher random effect estimates derived from a model using students' actual teachers and controlling for classroom peer characteristics (Column 3 of Table 6).

Appendix E

Table E1: Correlations of Teacher Effects on State Tests, Complex Tasks, and Social-Emotional Measures from Models Using Randomly Assigned Teachers

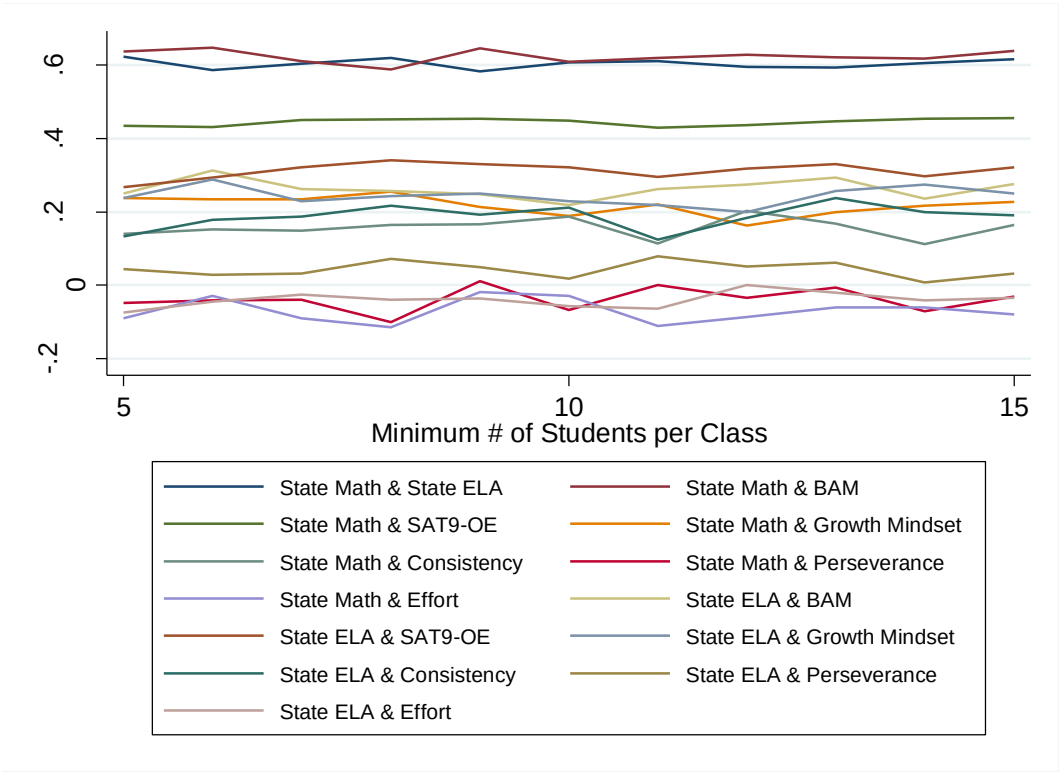
	State Math	State ELA	BAM Math	SAT9-OE Reading	Growth Mindset	Grit: Consistency	Grit: Perseverance
Panel A: Results from Models Using Randomly Assigned Teachers							
State ELA	0.55***						
BAM Math	0.49***	0.27***					
SAT9-OE Reading	0.35***	0.18**	0.47***				
Growth Mindset	0.17*	0.12	0.04	0.14*			
Grit: Consistency	0.07	0.10	0.00	-0.05	0.22***		
Grit: Perseverance	-0.07	-0.04	0.07	0.23**	-0.03	0.01	
Effort in Class	-0.03	0.06	0.07	0.12	-0.06	0.01	0.63***
Panel B: Results from Models Using Actual Teachers in Table 7							
State ELA	0.58***						
BAM Math	0.57***	0.31***					
SAT9-OE Reading	0.38***	0.24***	0.46***				
Growth Mindset	0.21***	0.12	0.10	0.19**			
Grit: Consistency	0.17*	0.21**	0.05	-0.03	0.22***		
Grit: Perseverance	-0.04	0.00	0.10	0.19**	-0.03	0.04	
Effort in Class	0.05	0.09	0.14*	0.10	-0.08	0.06	0.61***

Notes: *p<0.05; **p<0.01; ***p<0.001. n=229 in Panel A and n=227 in Panel B. This table reports unadjusted Pearson product-moment correlations using post-hoc predicted BLUE teacher random effect estimates derived from a model using students' randomly assigned teachers in Panel A, and actual teachers in Panel B, where both include controls for classroom peer characteristics (Column 5 of Table 6).

Appendix F

I examine the sensitivity of my results in Table 7 by re-estimating the teacher effects correlation matrix using a common subsample of teachers that have a minimum of 15 students in their class (between 96 and 104 teachers across pairwise combinations). I then repeatedly drop one student per teacher and re-estimate teacher effects and a corresponding correlation matrix until the minimum class size reaches five students. Figure C1 illustrates the relative stability of the estimated correlations as the sample size increases. These findings suggest the post-hoc predicted BLUE random effect estimates I use when correlating teacher effects sufficiently correct for sampling error among this limited range.

Panel A:



Panel B:

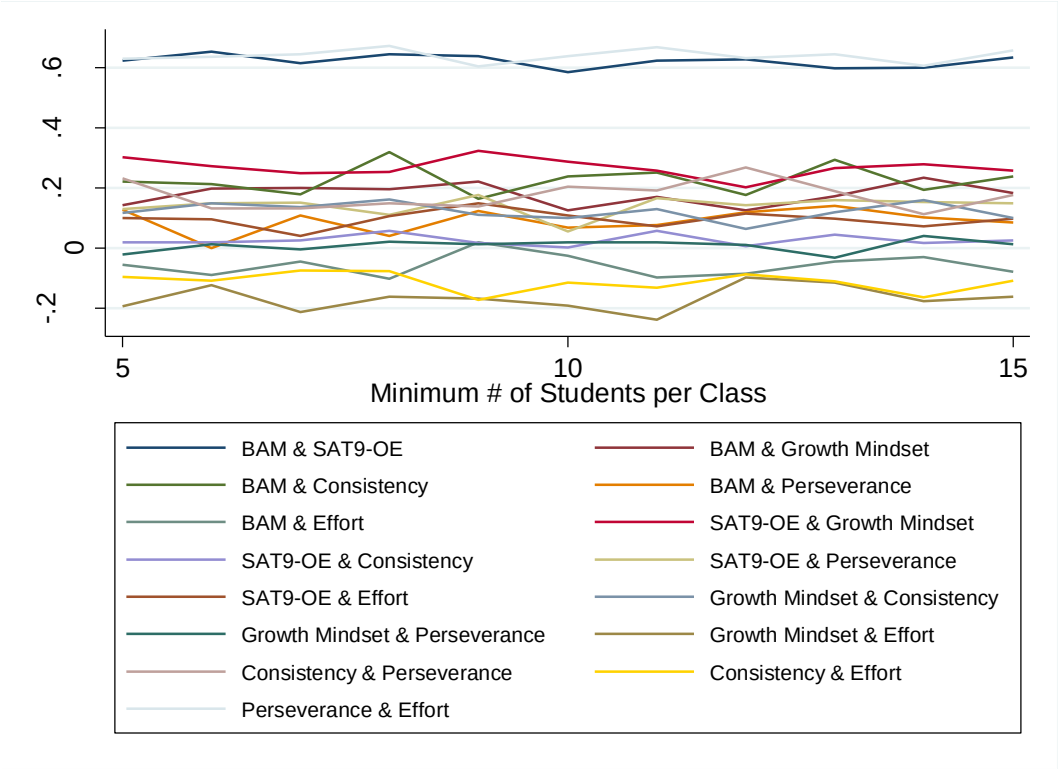


Figure F1: Trends in pairwise Person product-moment correlations of empirical Bayes teacher effect estimates from Table 7 across class size using successively larger minimum class size requirements. (N=96 to 104 teachers across pairwise combinations). Panel A includes correlations with state tests. Panel B includes correlations with measures of complex cognitive skills and social-emotional competencies.

Notes: Empirical Bayes estimates are Best Linear Unbiased Estimators of teacher random effects derived from the ML model that uses students' actual teachers and includes peer controls (Column 3 of Table 6).

Appendix G

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 Column 3 provides model-based ML estimates 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 random effects from ML models ($\overline{SE_{\tau_j}^2}$).

Table G1: Estimated Reliabilities of Teacher Effects

State Math	0.539
State ELA	0.592
BAM Math	0.561
SAT9-OE Reading	0.527
Growth Mindset	0.550
Grit: Consistency	0.481
Grit: Perseverance	0.544
Effort in Class	0.516

Notes: Reliabilities are estimated using sample analogues.

Table G2: Disattenuated Correlations among Teacher Effects on State Tests, Complex Tasks and Social-Emotional Measures

	State Math	State ELA	BAM Math	SAT9-OE Reading	Growth Mindset	Grit: Consistency	Grit: Perseverance
				Panel A: Disattenuated Correlations			
State ELA	1.00						
BAM Math	1.00	0.62					
SAT9-OE Reading	0.64	0.45	0.79				
Growth Mindset	0.42	0.33	0.22	0.41			
Grit: Consistency	0.35	0.37	0.19	-0.04	0.43		
Grit: Perseverance	-0.11	-0.04	0.18	0.34	-0.04	0.06	
Effort in Class	0.13	0.14	0.26	0.17	-0.09	0.12	1.00
				Panel B: Unadjusted Correlations			
State ELA	0.58***						
BAM Math	0.57***	0.31***					
SAT9-OE Reading	0.38***	0.24***	0.46***				
Growth Mindset	0.21***	0.12	0.10	0.19**			
Grit: Consistency	0.17*	0.21**	0.05	-0.03	0.22***		
Grit: Perseverance	-0.04	0.00	0.10	0.19**	-0.03	0.04	
Effort in Class	0.05	0.09	0.14*	0.10	-0.08	0.06	0.61***

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. Panel A reports disattenuated Pearson product-moment correlations using post-hoc predicted BLUE teacher random effect estimates derived from a model using students' actual teachers and controlling for classroom peer characteristics (Column 3 of Table 6). Correlations are adjusted using the Spearman (1904) correction for attenuation bias based on sample estimates of the reliability of each measure. Disattenuated correlation coefficients are set to 1 when they are exceed the possible range.

Table G3. Correlations of Teacher Effects Estimated by the MET Project Using a Covariate Adjustment Model.

	State Math	State ELA	BAM Math	SAT9-OE Reading
	Panel A: Same Class of Students			
State ELA	0.47			
BAM Math	0.38	0.28		
SAT9-OE Reading	0.23	0.27	0.35	
Effort in Class	0.18	0.15	0.11	0.10
	Panel B: Different Classes of Students Across Years			
State ELA	0.26			
BAM Math	0.17	0.16		
SAT9-OE Reading	0.06	0.06	0.16	
Effort in Class	0.12	0.09	0.00	0.02

Notes: n=236 teachers. Table reports unadjusted Pearson product-moment correlations. Panel A captures the pooled average correlation between teacher effects from the same class using data from 2010 and 2011. Panel B captures the pooled average correlation between teacher effects from classes in different years using estimates derived from combinations where measures in columns are from 2011 and rows are from 2010 and then vice versa. Teacher effects are estimated and provided by the MET Project using a standard covariate adjustment model using students' actual teachers and including all students taught by a teacher.

Appendix H

Table H1: Model-based Restricted Maximum Likelihood Estimates of Teacher Effects on State Tests, Complex Tasks and Social-Emotional Measures without Prior State Test Scores

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
		Panel A: Results from Models without Prior State Test Scores				
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*
		Panel B: Results from Models with Prior State Test Scores from Table 6				
State Math	4,075	0.175***	0.159***	0.150***	0.139***	0.124***
State ELA	4,074	0.142***	0.135***	0.137***	0.123***	0.123***
BAM Math	3,746	0.137***	0.143***	0.126***	0.129***	0.110**
SAT9-OE Reading	3,766	0.168***	0.177***	0.174***	0.176***	0.169***
Growth Mindset	3,551	0.196***	0.146**	0.133*	0.159***	0.154**
Grit: Consistency	3,473	0.082	0.077	0.078	0.080	0.102
Grit: Perseverance	3,473	0.149**	0.149**	0.136*	0.153**	0.138*
Effort in Class	3,435	0.162***	0.157**	0.183***	0.115*	0.149**
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 estimates of the standard deviation of teacher effects from separate regressions. Columns 1 through 3 estimate the effect of 4th and 5th grade students' actual teacher while columns 4 and 5 estimate intent-to-treat effects of the teachers students were randomly assigned to via the MET classroom roster randomization process. Panel A omits prior measures of student achievement on state standardized tests in math and reading while Panel B includes these measures. All models include controls for students' 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 H2: 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-OE Reading	Growth Mindset	Grit: Consistency	Grit: Perseverance
Panel A: Results from Models without Prior State Test Scores							
State ELA	0.73***						
BAM Math	0.73***	0.55***					
SAT9-OE Reading	0.47***	0.40***	0.55***				
Growth Mindset	0.27***	0.25***	0.20**	0.26***			
Grit: Consistency	0.37***	0.40***	0.29***	0.14*	0.27***		
Grit: Perseverance	0.07	0.13	0.20**	0.26***	0.02	0.13	
Effort in Class	0.17*	0.21**	0.23**	0.19**	-0.02	0.15*	0.63***
Panel B: Results from Models with Prior State Test Scores from Table 7							
State ELA	0.58***						
BAM Math	0.57***	0.31***					
SAT9-OE Reading	0.38***	0.24***	0.46***				
Growth Mindset	0.21***	0.12	0.10	0.19**			
Grit: Consistency	0.17*	0.21**	0.05	-0.03	0.22***		
Grit: Perseverance	-0.04	0.00	0.10	0.19**	-0.03	0.04	
Effort in Class	0.05	0.09	0.14*	0.10	-0.08	0.06	0.61***

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. This table reports unadjusted Pearson product-moment correlations using post-hoc predicted BLUE teacher random effect estimates derived from a model using students' actual teachers and controlling for classroom peer characteristics (Column 3 of Table 6).

Appendix I

Table I1: Unshrunken and Shrunken Average Residual 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)
		Panel A: Unshrunken Average Class Residuals				
State Math	4,075	0.171	0.150	0.131	0.141	0.122
State ELA	4,074	0.153	0.140	0.131	0.135	0.126
BAM Math	3,744	0.168	0.158	0.141	0.148	0.133
SAT9-OE Reading	3,766	0.208	0.197	0.177	0.197	0.178
Growth Mindset	3,551	0.263	0.220	0.194	0.228	0.209
Grit: Consistency	3,473	0.213	0.190	0.175	0.199	0.189
Grit: Perseverance	3,473	0.237	0.222	0.202	0.223	0.203
Effort in Class	3,435	0.250	0.228	0.212	0.208	0.202
		Panel B: Shrunken Average Class Residuals				
State Math	4,075	0.112	0.086	0.059	0.045	0.008
State ELA	4,074	0.078	0.062	0.050	0.011	0.023
BAM Math	3,744	0.077	0.065	0.034	0.002	0.060
SAT9-OE Reading	3,766	0.102	0.087	0.055	0.040	0.003
Growth Mindset	3,551	0.150	0.076	0.008	0.018	0.128
Grit: Consistency	3,473	0.032	0.025	0.000	0.000	0.000
Grit: Perseverance	3,473	0.072	0.051	0.001	0.000	0.000
Effort in Class	3,435	0.098	0.066	0.039	0.059	0.111
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 in panel A report estimates from separate models of the standard deviation of teacher effects estimated by averaging student-level residuals from an OLS model to the teacher level. Cells in panel B report these same estimates from separate models when shrunken towards the grand mean based on the reliability of a teacher's individual estimate. Statistical significance not calculated. All models include controls for students' prior achievement in math and reading, 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.