TEACHING FOR TOMORROW'S ECONOMY? TEACHER EFFECTS ON COMPLEX COGNITIVE SKILLS AND SOCIAL-EMOTIONAL COMPETENCIES

Matthew A. Kraft*
Sarah Grace
Brown University

September 2016

Abstract

We exploit the random assignment of class rosters in the Measures of Effective Teaching Project to estimate teacher effects on students' performance on cognitively demanding open-ended tasks in math and reading, as well as their growth mindset, grit, and effort in class. We find substantial variation in teacher effects on students' complex cognitive skills and social-emotional competencies, and weak relationships between these effects and traditional estimates based on state standardized tests. We show that high-stakes decisions based on existing teacher performance measures largely fail to consider their ability to develop skills and competencies highly-valued by employers.

JEL No. H0, I2, J24

* Correspondence regarding the paper can be sent to Matthew Kraft at mkraft@brown.edu. PO Box 1983, Brown University, Providence RI, 02912. This research was generously supported by the William T. Grant Foundation. The authors thank seminar participants at Brown, Harvard, Stanford, the University of Connecticut and the Federal Reserve Banks of Boston and New York for their helpful comments. Bruna Lee, Dylan Hogan, and Harry Neuert provided excellent research assistance in support of this paper. All views and errors are the authors' own.

1. Introduction

Students today are preparing to enter a labor market very different from that of their parents. Many of the middle-class jobs in which workers performed routine tasks have been replaced by automatized machines or moved offshore as a result of global market integration (Acemoglu and Autor 2011). Demand and wages for workers capable of performing more complex, non-routine tasks increased rapidly in the decades leading up to the turn of the century (Autor, Levy and Murnane 2003), but have since plateaued (Autor and Price 2013). Scholars have identified at least two factors that might explain this recent pattern – the growing supply of high-skilled workers (Autor 2014) and advances in machine learning that allow computers to substitute for jobs higher up in the skill distribution (Lu 2015).

The jobs that have proven most difficult to automate are those that require a combination of both high cognitive skills and so-called "soft" skills. For example, Deming (2015) and Weinberger (2014) find sustained employment growth among jobs with high cognitive demands that also require strong social skills, and Heckman and his colleagues have documented the joint importance of cognitive and "noncognitive" skills in the labor market (Heckman, Stixrud and Urzua 2006; Heckman and Kautz 2012). The challenge for the U.S. education system is to prepare students with both the increasingly complex set of cognitive skills as well as the social-emotional competencies required to succeed in the 21st century economy.

Scholars, policymakers and practitioners have voiced deep concern that the U.S. education system has not kept pace with the changing economic times (National Research Council 2012). These concerns about the alignment between educational practices and the evolving labor market in the U.S. are not new. In 1983, the landmark report *A Nation at Risk* warned that "schools may emphasize such rudiments as reading and computation at the expense

of other essential skills such as comprehension, analysis, solving problems, and drawing conclusions." In the decades since, the importance of complex cognitive skills as well as their complementarity with social-emotional competencies has only increased. In a 2014 survey of large commercial firms, corporate leaders identified both problem-solving skills and a strong work ethic as being among the most important attributes for job candidates (National Association of Colleges and Employers 2015). However, the degree to which teachers are developing students' abilities to apply knowledge in new contexts, learn on the job, and solve unstructured tasks through a combination of creativity, adaptability, and sustained effort remains an open question.

Although it is well established that teachers have large effects on student achievement, current evidence is largely limited to student performance on state standardized tests (Rockoff 2004; Hanushek and Rivkin 2010, Chetty, Friedman and Rockoff 2014a). These tests typically measure students' core content knowledge as well as basic literacy and numeracy skills using multiple-choice questions. Very few of these standardized exams include test items that assess students' performance on cognitively demanding open-ended tasks that provide more direct measures of the reasoning, inference and analytic skills required in the 21st century economy (Yuan and Le 2012). Using data from the Project STAR class-size experiment, Chetty and his colleagues (2011) estimate that only 1/5 of the total variation in "class effects" on earnings operates through effects on scores from multiple-choice tests.

The degree to which teachers are able to impact social-emotional competencies such as self-regulation and academic mindsets is even less understood. Are teachers developing these skills and competencies among their students? Are teachers who are successful at raising student achievement on state standardized tests the same as those who develop students' social-

emotional competencies? Do teacher performance measures capture their ability to develop these skills among their students? Answers to these questions have important implications for school accountability systems, teacher preparation programs, teacher performance evaluation systems and the future competitiveness of the U.S. workforce.

This paper provides new evidence to inform these questions using large-scale data across six large school districts collected by the Measures of Effective Teaching (MET) Project. In addition to collecting administrative and achievement data, MET researchers administered two supplemental achievement tests with open-ended questions that were designed to be more direct measures of students' critical thinking skills and problem-solving skills on open-ended tasks. In the second year of the study, students also completed a questionnaire that included scales for measuring their grit (Duckworth and Quinn 2009) and growth mindset (Dweck 2006), two widely-publicized social-emotional competencies that have received considerable attention from policymakers and the media in recent years.[1] The survey also included a class-specific measure of effort which allows us to compare teacher effects on both global and domain-specific social-emotional measures.

In the second year of the MET Project, a subset of teachers participated in an experiment where researchers randomly assigned student rosters among sets of volunteer teachers in the same grades and schools. The MET Project's experimental design, combined with its scale and unique set of measures, allow us to make several important contributions. Past MET Project reports have primarily focused on developing a composite measure of teacher effectiveness for

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

forecasting effects on student achievement (Kane and Staiger, 2012) and validating this measure using random assignment (Kane et al. 2013).[2] We extend this work by identifying "class effects" on students' ability to perform complex tasks as well as on students' grit, growth mindset and effort in class. We then present among the first direct evidence of the relationship between "class effects" on state tests, open-ended cognitively demanding assessments, and social-emotional competencies. Finally, we conduct simulation exercises to examine the degree to which high-stakes decisions made under new teacher evaluation systems incorporate information about teachers' effects across our broad range of student outcomes.

The nature of the MET Project randomization design required principals to create classroom rosters during the early summer and resulted in substantial attrition from the study. Critically, we find that this attrition was largely random within randomization blocks. Among those students randomized to teachers who appear in our data, approximately three out of four complied with their random assignment. We guard against the potential threat posed by non-compliance by conditioning on a rich set of student characteristics from administrative data and supplemental surveys, and by estimating intent-to-treat effects of students' randomly assigned teachers. We resolve potential concerns about reference bias in students' self-reported measures due to different normative expectations across schools or grades (West et al. 2016; Dobbie and Fryer 2013) with the inclusion of randomization blocks in our models which restrict variation to within the same schools and grades.[3] Exploratory tests for reference bias at the classroom level reveal little evidence of bias.

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

[3] For a detailed discussion of potential reference bias in self-report measures of social-emotional skills see Duckworth and Yeager (2015).

These analyses build on an emerging body of literature that examines teacher effects on student outcomes not captured by standardized tests (Blazar and Kraft 2015; Gershenson, 2016; Jackson 2016; Jennings and DiPrete 2010; Koedel 2008; Ladd and Sorensen, in press; Ruzek et al 2014). To date, these studies have primarily relied on a covariate adjustment approach with fixed effects to account for student sorting. Data from MET Project provide a unique opportunity to leverage the random assignment of class rosters to teachers to guard against the potential threats posed by student sorting in prior analyses. Our use of the MET data, however, also limit the type and scope of analyses we can perform in two important ways.

First, we focus our empirical analyses on the effect of general education elementary classrooms because we examine global measures of students' social-emotional skills which are not specific to any one class. This approach, along with the single year of experimental data limits our ability to isolate teacher effects from peers effects and transitory shocks. We refer to our "class effect" estimates as teacher effects throughout the paper while recognizing that the inclusion of a rich set of classroom-level controls does not definitively separate the joint effect of teachers, peers, and shocks.[4]

Second, we would ideally like to test the predictive validity of our estimated teacher effects on complex cognitive skills and social-emotional competencies using long-term outcomes. For example, Chetty et al. (2014b) documented the predictive power of estimated teacher effects on test scores for students' earnings. Particularly relevant to this work, Jackson (2016) found that teacher effects on an index of behavioral outcomes were predictive of high school graduation as well as predictors of college-going, above and beyond teacher effects on

[4] Blazar & Kraft (2015) directly examine the degree to which estimates that conflate teacher effects with peer effects and classroom shocks are biased upwards. They find that estimates of teacher effects that do not allow for class-specific effects are only slightly larger than estimates which simultaneously model both. Their findings suggest the degree to upward bias in our estimates is approximately 15%.

test scores. Similar analyses using the MET data are not possible because the MET Project focused on teachers and, thus, did not collect panel data on students. Given this limitation, we turn to 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.[5]

Using our randomization identification strategy, we find teacher effects on standardized achievement in math and English Language Arts (ELA) that are similar in magnitude to prior estimates (see Hanushek and Rivkin 2010). We also find teacher effects of comparable magnitude on students' ability to perform complex tasks in math and ELA, as measured by cognitively demanding open-ended tests. Teacher effects on students' social-emotional competencies differ in magnitude, with the largest effects on growth mindset, effort in class and the perseverance subscale of grit. Comparing the effects of individual teachers across these outcomes reveals that teachers who are most effective at raising student performance on standardized tests are not consistently the same teachers who develop students' complex cognitive abilities and social-emotional competencies. While teachers who add the most value to students' performance on state tests in math do also appear to strengthen their analytic and problem-solving skills, teacher effects on state ELA tests are only moderately correlated with teacher effects on open-ended tests in reading. Successfully teaching more basic reading comprehension skills does not appear to translate consistently to the ability to interpret and respond to texts. Furthermore, teacher effects on social-emotional measures are only weakly correlated with effects on state achievement tests and more cognitively demanding open-ended

[5] Evidence of the predictive validity of individual student measures for long-term outcomes, while suggestive, is not alone sufficient evidence to confirm the predicative validity of teacher effects on these measures.

tasks, even after adjusting for differential reliability in the measures. These findings suggest that teacher effectiveness differs across specific abilities.

We conclude by simulating high-stakes human capital decisions based on teacher evaluation scores constructed from classroom observations, teacher effects on state tests, principal ratings, and student surveys. Teachers identified as being in the top and bottom 15% of evaluation score rankings vary considerably in their effects on complex cognitive skills and social-emotional competencies. We show further that none of the individual teacher performance measures available in the MET data serve as reasonable proxies for the teacher effects we estimate. We discuss the implications of our 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 globalized economy (National Resource Council 2012; OECD 2013). Psychologists and learning scientists define complex cognitive skills as a set of highly interrelated constituent skills that support cognitively demanding processes (Van Merriënboer and Jeroen, 1997). These skills allow individuals to classify new problems into cognitive schema and then to transfer content and procedural knowledge from familiar schema to new challenges. Examples include writing computer programs, directing air traffic, engineering dynamic systems, or diagnosing sick patients.

Researchers and policy organizations have referred to these abilities using a variety of different terms including "21st Century Skills," "Deeper Learning," "Critical-Thinking" and "Higher-Order Thinking." State standardized achievement tests in mathematics and reading

rarely include items designed to assess these abilities. A review of standardized tests used in 17 states judged as having the most rigorous state assessments found that 98% of items on math tests and 78% of items on reading tests only required students to recall information and demonstrate basic skills and concepts (Yuan and Le 2012). Open-ended ELA questions on state tests were substantially more likely to be judged as cognitively demanding assessments of "deeper learning." However, while open-ended test items in math required students to move beyond recall, they rarely required students to perform extended unstructured problems.

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

2.2 Social-emotional Competencies

Social-emotional competencies (or social and emotional learning) is an umbrella term used to encompass an interrelated set of cognitive, affective and behavioral abilities that are not commonly captured by standardized tests. Although sometimes referred to as non-cognitive skills, personality traits, or character skills, these competencies explicitly require cognition, are not fixed traits, and are not intended to suggest a moral or religious valence. They are skills,

attitudes and mindsets which can be developed and shaped over time (Duckworth and Yeager 2015). Regardless of the term used, mounting evidence documents the strong predictive power of competencies other than performance on cognitive tests for educational, employment, health and civic outcomes (Almlund et al. 2011; Borghans et al. 2008; Moffitt et al. 2011). The salience of social-emotional competencies to important life outcomes has motivated researchers to examine the degree to which teachers and schools can develop these competencies in students (Cook, Murphy, and Hunt 2000; Heckman, Stixrud and Urzua 2006).

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).

Emerging evidence also suggests teachers can have large effects on their students' academic, behavioral, and social skills. Several studies have found evidence of teacher effects in primary (Gershenson 2016), middle (Ladd and Sorensen, in press), and high school (Jackson 2016; Koedel 2008) on non-test-based outcomes observable in administrative data such as absences, suspensions, grades, grade progression and graduation. Jennings and DiPrete (2010) found large kindergarten teacher effects on students' social and behavioral skills as assessed by their first grade teachers using the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K). Two additional studies have leveraged self-reported measures of students' attitudes

and behaviors captured on student surveys to estimates teacher effects. Blazar and Kraft (2015) found large teacher effects on students' self-efficacy in math, and behavior and happiness in class, while Ruzek et al. (2014) found smaller but statistically significant effects on students' motivation in math. Our study extends this literature by leveraging an experimental identification strategy and employing a range of previously unexamined measures.

3. Research Design

3.1 The MET Project

The MET Project was designed to evaluate the reliability and validity of a wide range of performance measures used to assess teachers' effectiveness. The study tracked approximately 3,000 teachers from across six large public school districts over the 2009-10 and 2010-11 school years.[6] These districts included the Charlotte-Mecklenburg Schools, the Dallas Independent Schools, the Denver Public Schools, the Hillsborough County Public Schools, the Memphis Public Schools, and the New York City Schools. Across districts there is substantial variation in the racial composition of students where African-American, Hispanic and white students each comprise the largest racial/ethnic group in at least one district.

In the second year of the study, MET researchers recruited schools and teachers to participate in a classroom roster randomized experiment. Of those 4th and 5th grade general education teachers who participated in the first year and remained in the study in the second year, 85% volunteered for the randomization study and were eligible to participate. Participating principals were asked to create classroom rosters that were "as alike as possible in terms of student composition" in the summer of 2010 (Bill and Melinda Gates Foundation 2013, p. 22).

[6] Detailed descriptions of the MET data are available at www.metproject.org.

They then provided these rosters to MET researchers to randomize among volunteer teachers in the same schools, subjects and grade levels.[7] 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.

We focus our empirical analyses on the effect of general education elementary classrooms to minimize the potential confounding caused when students are taught by multiple teachers. 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), we find substantial attrition among the 4th and 5th grade students who were included in the roster randomization process; 38.6% of students on these rosters were not taught by any teachers who participated in the MET Project data collection in 2010-2011 and thus are censored from our analyses. 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. We examine and discuss the implications of this attrition in section 5 and find that it does not compromise the internal validity of our analyses to a great degree.

We construct our analytic sample to include only students in 4th and 5th grades who 1) were included in the roster randomization process 2) were taught by general education teachers who participated in the randomization study, and 3) have valid lagged achievement data on state

[7] 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).

standardized tests in both math and ELA. These restrictions result in an analytic sample of 4,151 students and 236 general education teachers.

We present descriptive statistics on the students in our analytic sample in Table 1. Overall, our sample closely resembles the national population of students attending public schools in cities across the United States (Snyder and Dillow 2015): 36% are African-American, 29% are Hispanic, 24% are white, and 8% are Asian. Over 60% of students qualify for free or reduced-price lunch (FRPL) across the sample. We provide descriptive statistics for teachers in our analytic sample in Table 2. The 4th and 5th grade general education elementary school teachers who participated in the MET Project randomization design are overwhelmingly female. A majority of these teachers are white, while 33% are African American. Teacher experience varies widely across the sample, and half of all teachers hold a graduate degree.

3.2 Standardized State Tests

The MET dataset includes end-of-year achievement scores on state standardized tests in math and ELA, as well as scores from the previous year. Multiple-choice items were the primary question format used on the 4th and 5th grade state math and ELA tests administered in the six districts in 2011. State testing technical manuals suggest that the vast majority of items on these exams assessed students' content knowledge, fundamental reading comprehension and basic problem-solving skills.[8] Reported reliabilities for these 4th and 5th grade tests in 2011 range between 0.85-0.95. In order to make districts' scaled scores comparable across districts, the MET Project converted these scores into rank-based Z-scores.

[8] 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).

3.3 Achievement Tests Consisting of Open-Ended Tasks

MET researchers administered two supplemental achievement tests to examine the extent to which teachers promote high-level reasoning and problem solving skills. The cognitively demanding tests, the Balanced Assessment in Mathematics (BAM) and the Stanford Achievement Test 9 Open-ended Reading Assessment (SAT9-OE), consist exclusively of open-ended questions. The BAM was developed by researchers at the Harvard Graduate School of Education and consists of four to five tasks that require students to complete a series of open-ended questions about a complex mathematical problem and justify their thinking. Similar to the BAM, the SAT9-OE developed by Pearson Education consisted of nine open-ended questions about one extended reading passage that tested students' abilities to reason about the text, draw inferences, explain their thinking and justify their answers. We estimate internal consistency reliabilities of students' scores across individual items on the BAM and SAT-9-OE of 0.72 and 0.85, respectively.

3.4 Social-emotional Measures

Students completed short self-report questionnaires to measure their grit and growth mindset in the second year of the study. The scale used to measure grit was developed by Duckworth and Quinn (2009) 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*. Grit has been shown to be predictive of GPAs at an Ivy League school, retention at West Point, and performance in the Scripps National Spelling Bee, conditional on IQ (Duckworth et al. 2007; Duckworth and Quinn 2009). Grittier soldiers were more likely to complete an Army Special Operations Forces selection course, grittier sales employees were

more likely to keep their jobs, and grittier students were more likely to graduate from high school, conditional on a range of covariates (Eskreis-Winkler et al. 2014). Following Duckworth and Quinn (2009), we estimate student scores for two subscales of the grit measure, 1) consistency of interest and 2) perseverance of effort (hereafter consistency and perseverance).

The growth mindset scale developed by Dweck (2006) stems from decades-long work on implicit theories of intelligence. Growth mindset 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. 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*. Middle school students who report having a high growth mindset have been found to have higher rates of math test score growth than students who view intelligence as fixed (Blackwell, Trzesniewski and Dweck 2007). Two 45-minute computer models designed to increase high-school students' growth-mindset and sense-of-purpose have been shown to increase course passing rates among students at-risk of dropping out (Paunesku, et al 2015).

We 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 0.77, 0.65, 0.71, and 0.58 respectively. We construct scores on each of the measures following Duckworth and Quinn (2009) and Blackwell et al. (2007) by

assigning point values to the Likert-scale responses and averaging across the items in each scale. We 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 On-line Appendix A for the complete list of items included in each scale.

3.5 Predictive Validity of Grit and Growth Mindset

While a growing body of evidence documents that measures such as the Big Five, locus of control, and self-esteem are strong predictors of long-term life outcomes (Almlund et al. 2011; Borghans et al. 2008; Moffitt et al. 2011), evidence of the predictive validity of grit and growth mindset is limited to more immediate outcomes. Given the lack of medium or long term outcomes in the MET data, we conduct a simple validation exercise similar to Jackson (2016) using the Educational Longitudinal Study (ELS). Specifically, we 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. As predictors we use proxy measures of grit and growth mindset constructed from 10^{th} grade students' self-reported answers to survey items that map closely onto the perseverance of effort subscale of grit and provide a domain-specific measure of students' growth mindset in math. We create a composite measure of students' academic ability in math and reading based on students' scores on a multiple-choice achievement test administered by National Center for Education Statics (See On-line Appendix B for specific details on measures). We standardize all predictors so the magnitude of their coefficients can be interpreted using the same scale.

In Table 3 we report results from a simple set of OLS regressions which include controls for academic achievement, students' race and gender, and level of parental education and

household income. We find that our measures of grit and growth mindset are meaningful predictors of all four types of adult outcomes even when included simultaneously. 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. These social-emotional measures are also strong predictors of civic outcomes. Among students with similar achievement levels, a one standard deviation increase in grit and growth mindset is associated with 3.5 and 1.9 percentage point increases in the probability a student voted in the most recent presidential election.

These analyses suggest that among 10th grade students, self-reported measures of students' grit and growth mindset capture important skills that are distinct from academic ability. The relationships we find are of meaningful economic magnitudes, ranging from approximately one half to twice the predictive power of academic achievement. Further, our math-specific measure of growth mindset is likely a weaker predictor of adult outcomes than the global growth mindset measure used in the MET Project. Prior studies suggest that any reference bias in self-reported social-emotional measures would cause our estimates to understate the true conditional relationships between these measures and long-term outcomes (West et al. 2016; Dobbie and Fryer 2013).

3.6 Estimating the Variance of Teacher Effects

We begin by specifying an education production function to estimate teacher effects on student outcomes. A large body of literature has examined the consequences of different model specifications (Todd and Wolpin 2003; Kane and Staiger 2008; Koedel and Betts 2011; Guarino,

Reckase, and Wooldridge 2015; Chetty, Friedman, and Rockoff 2014a). Typically, researchers exploit panel data with repeated measures of student achievement to mitigate against student sorting by controlling for prior achievement. The core assumption of this approach is that a prior measure of achievement is a sufficient summary statistic for all the individual, family, neighborhood, and school inputs into a student's achievement up to that time. Models also commonly include a vector of student characteristics, averages of these characteristics and prior achievement at the classroom level, and school fixed effects (see Hanushek and Rivkin 2010).

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

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

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

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

Following Nye et al. (2004) and Chetty et al. (2011), we estimate the magnitude of the variance of teacher effects using a direct, model-based estimate derived via restricted maximum likelihood estimation. We assume a Gaussian data generating process which appears well justified in our data for state and open-ended tests and an appropriate approximation for social-emotional measures. This approach produces both a maximally efficient and consistent estimator for the true variance of teacher effects. To arrive at this model-based estimate, we specify a multi-level covariate-adjustment model as follows:

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

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

Here, the outcome of interest, Y_{ij}, is a given outcome for student *i* in district *d*, in grade *g*, with teacher *j* in school *s* in year *t*. We model each outcome as a cubic function of students' prior year achievement on state standardized tests ($A_{i,t-1}$), in both mathematics and ELA. We allow the relationship between prior test scores and outcomes to vary across districts and grades by interacting the linear lagged test score terms with district-by-grade fixed effects. Additional covariates include the average prior achievement in a student's class in both subjects ($\bar{A}_{j,t-1}$); a vector of controls for observable student characteristics (X_i); average student characteristics in a students' class ($\bar{X}_j$); and randomization block fixed effects (π_{sg}).[9]

We 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. We obtain an estimate of the

[9] We include grade and school subscripts on randomization block fixed effects because these blocks are analogous to grade-by-school fixed effects.

true variance parameter, $\hat{\sigma}_{\tau}^{2}$, directly from the model through maximum likelihood estimation of this multilevel model. We specify τ_j in two different ways across models – with students' actual teachers and their randomly assigned teachers. Modeling the effects of students' actual teachers may lead to potentially biased estimates due to noncompliance with random assignment. Among those students in our analytic sample, 28.1% are observed with non-randomly assigned teachers. We address this 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).

We attempt to remove peer effects by controlling for a rich set of average classroom covariates.[10] When data are available for teachers who teach multiple classes in a given year or are observed across multiple years, researchers can model classroom-specific shocks due to peer effects via class or year random effects nested within teachers. Such an approach is unavailable to us; thus, we cannot definitely separate peer effects from teacher effects. Research suggests that, if anything, this approach produces conservative estimates of the magnitude of teacher effects (Kane et al. 2013; Thompson, Guarino, and Wooldridge 2015).

Across all models, student characteristics include indicators that control for a student's gender, age, race, FRPL, English proficiency status, special education status, and participation in a gifted and talented program.[11] We 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

[10] We calculate peer characteristics based on all students who were observed in a teacher's classroom, regardless of whether they were included in the classroom roster randomization process or not. In our view, this provide a more representative estimate of the peer environment in a classroom.

[11] Data on FRPL was not provided by one district. We account for this by including a set of district-specific indicators for FRPL and imputing all missing data as zero.

to which English is spoken at home, and the number of computers in their homes.[12] Both theory and prior empirical evidence have shown that grades reflect students' cognitive skills and as well as social-emotional competencies such as grit and effort (Bowen, Chingos, McPherson 2009). We find that our measure of grades is positively correlated with our social-emotional measures even when controlling for prior achievement in math and ELA. Partial correlations in our analytic sample range from 0.04 with growth mindset to 0.22 with perseverance. Classroom covariates include the means of all of these administrative and survey-based predictors. We restrict our estimation samples to exclude any classrooms where less than five students had valid outcome measures.

4. Findings

4.1 Achievement tests, performance on open-ended tasks, and social-emotional competencies

Correlations across our student outcomes of interest reveal a number of patterns about the relationship between achievement tests, performance on open-ended tasks, and social-emotional measures. In Table 4, we present Pearson correlations among our outcome measures. We see the strongest relationship among all eight outcomes is between students' performance on state standardized tests across subjects (0.74). Students' performance on the state tests and open-ended tests in math are also strongly related (0.66) suggesting 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 moderately correlated at 0.49, suggesting that state ELA tests are imperfect proxies for students' more complex reasoning and writing skills. Correlations between social-emotional measures and state tests as

[12] We impute values of zero for students with missing survey data and include an indicator for missingness.

well as open-ended tests are positive but of lower magnitude, ranging between 0.15 and 0.31. Relatively moderate correlations among the social-emotional measures themselves suggest that these scales capture two distinct competencies rather than a single latent construct, measures of 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. Disattenuating the correlations in Table 4 to account for measurement error increases their strength only moderately, while the relative magnitudes of the pairwise correlations remain largely the same (see On-line Appendix C for disattenuated correlations).

4.2 Post-Attrition Balance Tests

We conduct two tests to assess the degree to which student attrition from the original randomized classroom rosters poses a threat to the randomization design. We begin by testing for balance in students' average characteristics and prior achievement across classrooms in our analytic sample. We do this by fitting a series of models where we 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 5, we report F-statistics and associated p-values from a joint F-test of the significance of our full set of randomly assigned teacher fixed effects. We find that, post-attrition, students' characteristics and prior achievement remain largely balanced within randomization blocks. For ten of our twelve measures, such as race, special education status, eligibility for free or reduced-price lunch and prior achievement in math and ELA, we cannot reject the null hypothesis that there are no differences in average student characteristics across randomly assigned teachers. However, we do find evidence of imbalance in the proportion of students in randomly assigned teachers' classrooms based on

whether a student participated in a gifted program or was 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.

We 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 6, we present results from a series of regression models in which we regress individual student characteristics and prior achievement on prior-year value-added scores of their randomly assigned teachers. We 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.[13] Among the 48 different relationships we test, we find that only one is statistically significant at the 5% level. Post-attrition, teachers with a one standard deviation higher value-added score on the state math exam in the prior year are 2.8 percentage points more likely to be assigned a student who is from a low-income family. Furthermore, 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. Although a joint significance test does reject the null hypothesis of no relationship between students' characteristics and their randomly assigned teachers' value-added on state math exams, we find no overall relationship between observable student measures and teachers' value-added derived from ELA, BAM and SAT9-OE tests.

Together, these tests of post-attrition randomization balance across teachers suggest that the classroom roster randomization process did largely eliminate the large and systematic sorting

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

of students to teachers commonly present in observational data (Kalogrides and Loeb 2013; Rothstein 2010). Although we observe some differential attrition across classrooms based on students' gifted and ELL status, there is little evidence that this attrition is related to teachers' effectiveness. Neither gifted nor ELL students were more likely to remain in our sample if they were assigned to a teacher with higher prior value-added scores across four different tests. However, we cannot rule out the possibility that student attrition is related to the value-added of their randomly assigned teachers on state math tests. We attempt to address any potential threat to the randomization design posed by dynamic differential attrition by conditioning on a rich set of student and peer-level covariates.

4.3 Teacher Effects – Maximum Likelihood Estimates

In Table 7, we present estimates of the standard deviation of teacher effects from a range of models. We begin with a baseline model in Column 1 that corresponds to the predominant school fixed effect specification in the teacher effects literature reviewed by Hanushek and Rivkin (2010). Consistent with prior studies, our maximum likelihood estimates of the magnitude of teacher effects on state test scores are 0.16 standard deviations (sd) in math and 0.14 sd in ELA. Using this baseline model, we also find teacher effects on the BAM and SAT9-OE tests of 0.14 sd and 0.16 sd, respectively. Finally, we find suggestive evidence of teacher effects on social-emotional measures ranging from 0.09 sd for consistency of interest (not statistically significant) to 0.20 for growth mindset.

We present results from our preferred models where we exchange school fixed effects for randomization-block fixed effects. Across these models, we 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 we estimate teacher effects using students' actual teachers. In Columns 4 and 5, we exchange students' actual teachers with their randomly assigned teachers to arrive at an ITT estimate of the magnitude of teacher effects. For both specifications, we present results with and without peer effects to illustrate the degree to which peer controls attenuate our estimates. Comparing results across Columns 2 vs. 3 and 4 vs. 5 illustrates how the inclusion of peer-level controls somewhat attenuates our estimates by absorbing peer effects that were otherwise attributed to teachers. Focusing on estimates with students' actual teachers that condition on peer controls (Column 3), we find relatively similar estimates of the magnitude of teacher effects on most outcomes as in our baseline model. Teacher effects on growth mindset are attenuated (0.14 sd) and become similar in magnitude to effects on state tests.

As is common in field experiments in schools, there were some students who did not comply with the experimental design. In order to account for this non-compliance we estimate intent-to-treat effects of students' randomly assigned teachers. Results from these models are slightly attenuated given this non-compliance but remain consistent with estimates reported above. Teacher effects on academic outcomes range from 0.11 sd on the BAM to 0.16 sd for the SAT9-OE. Teacher effects on consistency of interest do not achieve statistical significance, while effects on students' growth mindset (0.16 sd), perseverance (0.14) and effort in class (0.14) are of similar and even slightly larger magnitude than effects on achievement. Together, these results present strong evidence of meaningful teacher effects on students' social-emotional competencies and ability to perform complex tasks.

4.4 Comparing Teacher Effects across Outcomes

We investigate the nature of teacher skills by examining the relationships between individual teacher's effects across our eight outcomes of interest. Pairwise correlations are

imperfect estimates of the true relationships between teacher effects on different outcomes, although the direction of potential bias is unclear. Because our estimates are derived from the same sample of students for each general education teacher, unobserved student traits correlated with multiple outcomes will likely induce an upward bias. At the same time, sampling error due to the small number of students per teacher and the imperfect reliability of our outcome measures will bias our estimates downward. We report results from unadjusted correlations below and examine the direction and magnitude of any biases in On-line Appendix D. Analyses suggest that, if anything, our estimates may be moderately biased downward but that even the most extreme corrections for this bias do not change the qualitative nature of our findings.

In Table 8, we present Pearson correlations of Best Linear Unbiased Predictor estimates of teacher effects from our maximum likelihood (ML) model that uses students' actual teachers and includes peer controls (Column 3 of Table 7).[14] Consistent with past research, we find that the correlation between general education elementary teachers' value-added on state math and ELA tests is large at 0.60 (Corcoran, Jennings and Beveridge 2012). Elementary teacher effects on state math tests also appear to be strongly related to their effects on the BAM (0.66). This suggests that teachers who are effective at teaching more basic computation and numeracy skills also appear to be developing their students' complex problem-solving skills. In contrast, teacher effects on state ELA exams are a poor proxy for teacher effects on more cognitively demanding open-ended ELA tests (0.25). In fact, teachers' value-added to students achievement on the SAT9-OE, which captures students' ability to reason about and respond to an extended passage, is most strongly related to their effects on the similarly demanding open-ended math test (0.43).

We find that teacher effects on social-emotional measures are only weakly correlated

[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.

with effects on both state standardized exams and exams testing students' performance on complex tasks. Among the four social-emotional measures, growth mindset has the strongest and most consistent relationship with teacher effects on state tests and complex task performance, with correlations ranging between 0.12 and 0.23. Teachers' ability to motivate their students' perseverance and effort is consistently a stronger predictor of teacher effects on students' complex task performance than on standardized tests scores. Finally, we note that teacher effects across different social-emotional measures are far less correlated than teacher effects on student achievement across subjects. Teacher effects on growth mindset are positively correlated with effects on students' consistency of interest (0.22), but unrelated to a teacher's ability to motivate students' perseverance and effort. Teacher effects on perseverance and effort in class are the only two social-emotional measures that appear to be capturing the same underlying ability among teachers, with a correlation of 0.61. This is important because it suggests that teacher effects on students' willingness to devote effort to their classwork may extend to other contexts as well.

We illustrate the substantial degree of variation in individual teacher effects across measures by providing a scatterplot of teacher effects on state math tests and growth mindset in Figure 1. This relationship captures the strongest correlation across teacher effects on social-emotional competencies and state tests (0.23). A total of 43% of teachers in our sample have above average effects on one outcome but below average effects on the other (24% in quadrant II and 19% in quadrant IV). Only 31% of teachers have effects that are above average for both state math tests and growth mindset. The proportion of teachers who have above average effects on both state tests and other social-emotional measures is even lower.

4.5 Do Teacher Evaluation Ratings Reflect Teacher Effects on Complex Cognitive Skills and Social-emotional Competencies?

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

The MET Project provides a unique opportunity to further explore the degree to new teacher evaluation systems capture teachers' ability to impact students' complex cognitive skills and social-emotional competencies. We begin by examining the correlation between a range of evaluation measures and the teacher effects we estimate above. We utilize evaluation ratings on two widely used classroom observation instruments: the Framework for Teaching (FFT) and the Classroom Assessment Scoring System (CLASS) (Kane and Staiger 2012). We also include principals' overall ratings of teachers' performance on a six-point scale and students' opinions of their teachers' instruction captured on the TRIPOD survey (Kane and Cantrell 2010).

Table 9 provides correlations between evaluation measures and our estimated teacher effects. We find that neither observation scores, principal ratings, nor student surveys serve as close proxies. In fact, teacher effects on state standardized tests have the strongest correlation with teacher effects on complex cognitive skills in both subjects as well as growth mindset and consistency as reported in Table 8. These "value-added" measures, however, are only available for a fraction of teachers who teach in tested grades and subjects. Student surveys have the strongest relationship with teacher effects on students' perseverance and effort in class, even

when compared with teachers' "value-added" scores, but they are only correlated at 0.19. The closest proxy for teacher effects on complex tasks, growth mindset and consistency of interest outside of effects on state tests are observational scores, particularly on the FFT instrument although these correlations are never larger than 0.13.

We next illustrate how current high-stakes decisions based on teacher evaluation scores might be viewed in light of additional information about teacher effects on complex cognitive skills and social-emotional competencies. In Washington D.C. Public Schools (DCPS), for example, teachers who receive low performance ratings face dismissal, while those who are rated as highly effective can earn large bonuses and permanent salary increases (Dee and Wyckoff 2015). We begin by constructing a composite evaluation score for each teacher as a weighted linear sum of four evaluation measures, where weights reflect a prototypical evaluation system for teachers in tested grades and subjects (Steinberg and Donaldson, in press).[15]

(4) $Score = .55 * FFT + .35 * ValueAdded + .05 * Principal\ Rating + .05 * Survey$

Next, we rank order teachers based on their evaluation score and identify those teachers in the bottom and top 15% of the distribution – percentages that reflect the proportion of teachers in DCPS that face dismissal threats or earn bonus pay. In Figure 2, we plot the cumulative densities of these low and high rated teachers using composite measures of teacher effects on complex cognitive skills and social-emotional competencies.[16]

Figure 2 illustrates how the ability to develop students' complex cognitive skills and

[15] We standardize all four performance measures to be mean zero and have a variance of one.
[16] Composites are simple averages of teacher effects on the two measures of complex cognitive skills and four measures of social-emotional competencies. Results using the CLASS observation instrument are nearly identical.

social-emotional competencies varies widely among teachers in the top and bottom 15% of evaluation score rankings. We find that half of the teachers in the top 15% of evaluation ratings are ranked below the 78th percentile on complex task effects and half are ranked below the 67th percentile for social-emotional effects. Conversely, half of the teachers in the bottom 15% of evaluation ratings are ranked above the 40th percentile for complex task effects and half are ranked above the 47th percentile for social-emotional effects. Alternative weighting schemes produce very similar results. These findings suggest that high-stakes decisions based on common measures of teacher performance largely fail to consider the degree to which teachers are developing skills and abilities that have been shown to be critical for the future labor market.

5. Robustness Tests

5.1 Teacher Effects – Average Class Residual Estimates

As a robustness check for our preferred model-based maximum likelihood estimation approach, we also estimate the variance of teacher effects by averaging upper and lower bound estimates derived from a two-step estimation approach following Kane et al. (2013). This allows us to relax the random effects normality assumption necessary for equation (3). Given that teacher fixed effects are perfectly collinear with classroom-level controls in our analytic sample, we first fit the covariate-adjustment model described in equation (3), omitting teacher random effects. In a second step, we average student residuals at the teacher level, $\bar{\varepsilon}_{ij}$, to estimate teacher effects. The variance of these average classroom residuals provide an upper bound estimate. We then shrink our average classroom residuals as described in equation (1). Following Jacob and Lefgren (2008), we estimate λ_j using sample analogs where σ_τ^2 is approximated by subtracting the average of the squared standard errors of our average classroom residuals from

the variance of these average classroom residuals ($\hat{\sigma}^2_{\bar{\varepsilon}_{ij}} - \overline{SE^2_{\bar{\varepsilon}_{ij}}}$) and $\sigma^2_{\varepsilon_j}$ is the squared standard error of teacher *j*'s average classroom residuals ($SE^2_{\bar{\varepsilon}_{ij}}$).[17] The variance of these shrunken EB estimates provides a lower-bound estimate. Finally, we average our upper and lower bound estimates to approximate the true teacher variance.

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

Two broad findings emerge from comparing our alternative estimates in Table 10 to our preferred ML results in Table 7. First, the relative magnitude of teacher effects across outcomes remains similar to our ML estimates across model specifications. Second, the magnitudes of our alternative results are slightly smaller than our ML results. This attenuation is largely a mechanical product of the two-stage estimation approach. ML variance estimates are derived from models that include peer controls and teacher random effects simultaneously. In our two-stage process of estimating average class residuals, we first estimate peer effects and then use only the remaining residual variation to quantify teacher effects. If peer effects and teacher effects are correlated, this two-stage approach will cause some variation attributable to teachers to be removed via peer controls in the first stage.

5.2 Falsification Tests

At their core, our 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 our analytic sample—it is possible that our estimates are the result of sampling error across classrooms. We test for this by generating a random variable from the standard normal distribution so that it shares the same mean and variance as our

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

outcomes. We then re-estimate our taxonomy of models using these random values as our outcomes and report the results in Table 11. This test fails to reject the null hypothesis of no teacher effects, demonstrating that our estimates are not driven by small sample error.

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

5.3 Potential Reference Bias in Social-emotional Measures

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

We examine whether students' responses on self-report measures of grit, growth mindset and effort in class may be subject to reference bias in our sample of traditional public schools in large urban districts. We do this by exploring how the direction and magnitude of the relationship between these social-emotional measures and student achievement gains on state standardized tests change when collapsed from the student-level to the class- and school-levels. Employing this same test, West et al. (2016) find suggestive evidence of reference bias in self-reported measures of grit, conscientiousness and self-control in a sample students attending traditional, charter and exam schools in Boston. They find that correlations between social-emotional measures and overall student gains become negative when collapsed to the school-level. This is analogous to the classic example of reference bias in cross-cultural surveys where, despite a widely acknowledged cultural emphasis on conscientious behavior, individuals in East Asian countries rate themselves lower in conscientiousness than do individuals in any other region (Schmitt et al. 2007).

We find no compelling evidence of reference bias at either the class level or the school level in the MET data. As shown in Table 12, simple Pearson correlation coefficients between our 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 we cannot rule out the potential of reference bias in our measures, it does not appear as though teachers or schools where students are making larger achievement gains are also systematically changing students' perceptions of what constitutes gritty behavior and high levels of effort. Additionally, our experimental design limits our identifying variation to

within school-grade cells, eliminating any potential for reference bias at the school-level and grade-level within a school.

6. Conclusion

Structural changes in the U.S. economy are placing increased pressure on schools to prepare students with a broader and more complex set of fundamental skills than the traditional domains of reading, writing and arithmetic. These skills include the ability to acquire new knowledge, to apply knowledge in new contexts, to interpret and respond to text, to solve unstructured problems, to sustain effort towards long-term goals, to persevere when faced with challenges, and to work productively in teams. However, the hallmark education policy reforms of the 21st century—accountability systems and teacher evaluations—have focused narrowly on measures of students' core content knowledge and basic skills. Questions remain about whether those teachers and schools that are judged as effective by state standardized tests are also developing the skills necessary to succeed in the 21st century economy. Our results suggest that this is not always the case. The lack of medium to longer-term student outcomes in the MET Project point to the importance of future studies that can directly examine the link between teacher effects on complex cognitive skills and social-emotional competencies, and students' life-outcomes.

The large differences in teachers' ability to raise student performance on achievement tests (Chetty, Friedman and Rockoff 2014a; Hanushek and Rivkin 2010) and the inequitable distribution of those teachers who are most successful at raising achievement (Clotfelter, Ladd, Vigdor 2006; Lankford, Loeb, Wyckoff 2002) have become major foci of academic research and education policy. The substantial variation we find in teacher effects on students' complex task

performance and social-emotional competencies further reinforces the importance of teacher quality but complicates its definition. Measures of teachers' contribution to their students' performance on state tests in math are strong proxies for their effects on students' ability to solve complex math problems. However, teacher effects on state ELA tests contain more limited information about how well a teacher is developing students' abilities to reason about and draw inferences from texts. Teacher effects on state tests are even weaker indicators of the degree to which teachers are developing students' social-emotional competencies. Even teachers who excel at developing competencies such as grit are not consistently the same as those that develop other competencies such as growth mindset.

Teaching core academic skills along with social-emotional competencies and the ability to perform unstructured tasks should not be viewed as competing priorities in a zero sum game. Elevating the importance of these new foundational skills does not require schools to make tradeoffs such as deciding between expanding instructional time in core subjects or teaching the arts and foreign languages. Our data suggest that there are teachers who teach core academic subjects in ways that also develop students' complex problem-solving skills and social-emotional competencies. We need to know what instructional practices allow these teachers to develop a wider range of students' skills and competencies than are commonly assessed on state achievement tests.

Current accountability and evaluation systems in education provide little incentive for teachers to focus on helping students develop the complex problem-solving skills and social-emotional competencies required in the fastest growing sectors of the economy. Our findings suggest that neither observation scores, principal ratings, nor student surveys are serving as close proxies for teacher effects on these skills and competencies. New computer-adaptive assessments

aligned with the Common Core State Standards move in this direction but are facing growing opposition. Developing reliable measures of students' social-emotional competencies poses an even greater challenge. Psychologists have argued that the social-emotional measures used in this study are not sufficiently robust to be used in high-stakes settings to compare teachers across schools (Duckworth and Yeager 2015). Student self-reports or teacher assessments of social-emotional measures are easy to game, and we know little about their properties when stakes are attached. There exists real potential to improve the reliability and robustness of these measures, although, as Einstein observed, "Everything that counts cannot necessarily be counted."

Preparing students for success in the new economy will require not only realigning incentives in the education sector, but also building teachers' capacity to develop students' complex cognitive skills and social-emotional competencies. Most teacher education and professional development programs provide little guidance on specific pedagogical approaches for developing these skills and competencies. Identifying and integrating these instructional approaches into common professional practice in K-12 classrooms will be a challenging but worthwhile investment in students' futures and our economy.

References

Acemoglu, Daron, and David Autor. "Skills, Tasks and Technologies: Implications for Employment and Earnings," *Handbook of labor economics*, 4 (2011), 1043-1171.

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

Autor, David, "Polanyi's Paradox and the Shape of Employment Growth," NBER Working Paper No. w20485, National Bureau of Economic Research, 2014.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics*, 118 (2003), 1279-1333.

Autor, David.H., and Brendan Price, "The Changing task Composition of the US Labor Market: An Update of Autor, Levy, and Murnane (2003)." MIT Working Paper, 2013. http://economics.mit.edu/files/9758

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

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

Blazar, David., Matthew A. Kraft, "Teacher and Teaching Effects on Students' Academic Behaviors and Mindsets." Mathematica Policy Research Working Paper. (2015). http://www.mathematica-mpr.com/our-publications-and-findings/publications/teacher-and-teaching-effects-on-students-academic-behaviors-and-mindsets

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

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

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

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

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

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

Cook, Thomas D., Robert F. Murphy, and H. David Hunt, "Comer's School Development Program in Chicago: A Theory-Based Evaluation." *American Educational Research Journal*, 37 (2000), 535-597.

Corcoran, Sean. P., Jennifer L. Jennings, and Andrew A. Beveridge, "Teacher Effectiveness on High-and Low-Stakes Tests." New York University Working Paper, 2012. http://www.nyu.edu/projects/corcoran/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf

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

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

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

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

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

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

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

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

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

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

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

Heckman, James J., and Tim Kautz, "Hard Evidence on Soft Skills," *Labour Economics*, 19 (2012), 451-464.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

National Association of Colleges and Employers. Job outlook 2015. (2014). https://www.naceweb.org/surveys/job-outlook.aspx

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Zellner, Arnold, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57 (1962), 348-368.

Figures

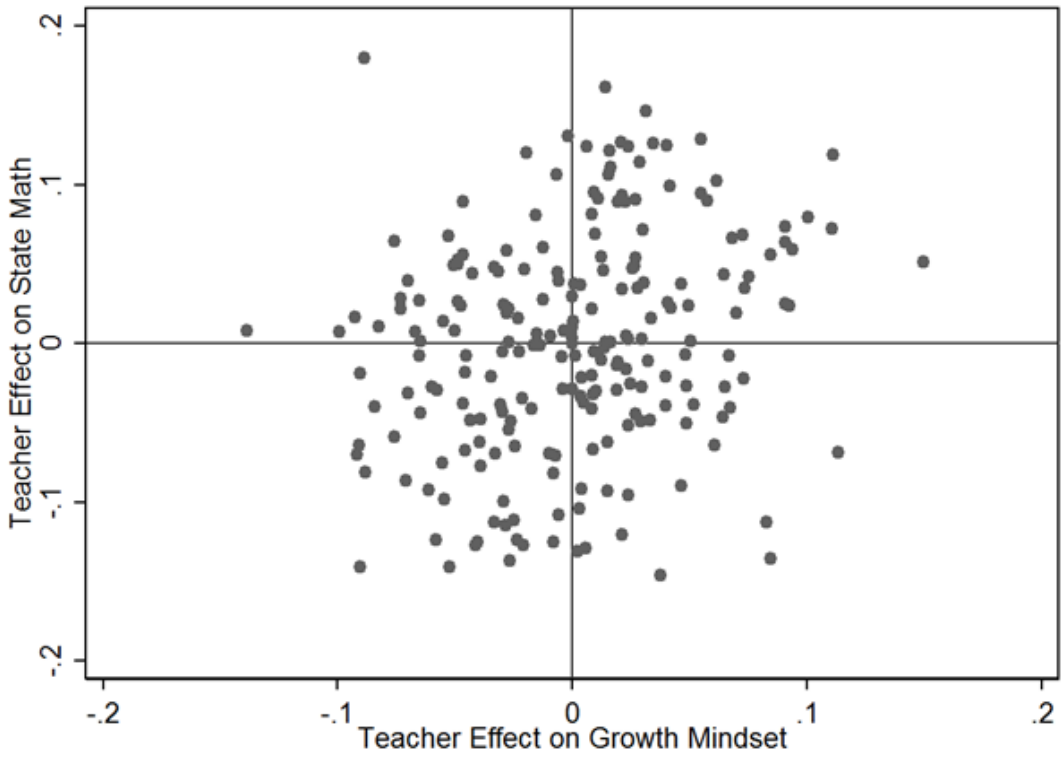


Figure 1: Scatterplot of teacher effects on state math test and growth mindset.

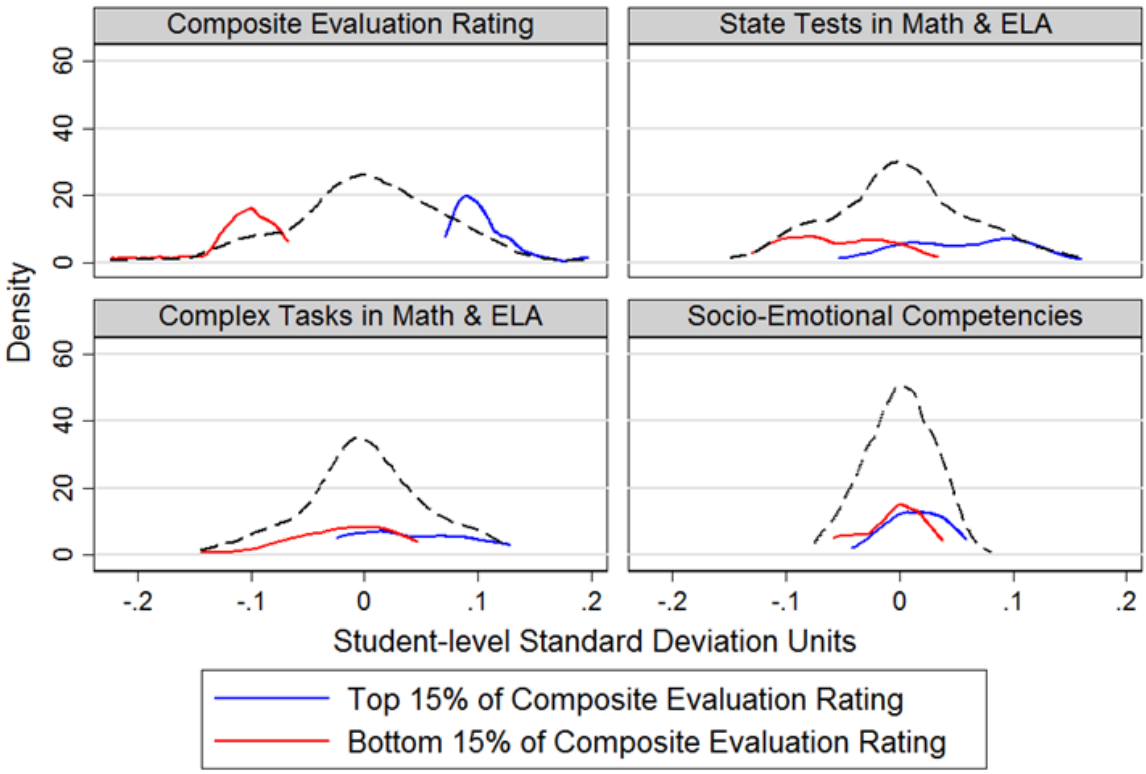


Figure 2: Density plots of teachers in the top and bottom 15% of evaluation ratings across multiple measures of teacher effectiveness

Tables

Table 1: Student Characteristics

Male	0.49
Age	9.50
Gifted Status	0.07
Special Education Status	0.08
English Language Learner	0.15
Free or Reduced Price Lunch	0.62
White	0.24
African American	0.36
Hispanic	0.29
Asian	0.08
Fourth Grade	0.49
Fifth Grade	0.51
n students	4092

Notes: The 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 socio-emotional outcome, as well as prior test scores on both math and ELA state exams.

Table 2: Teacher Characteristics

Male	0.08
White	0.62
African American	0.33
Hispanic	0.05
Other Race	0.01
1 Year of Experience in District	0.07
2-3 Years of Experience in District	0.18
4-6 Years of Experience in District	0.23
7-10 Years of Experience in District	0.24
11-20 Years of Experience in District	0.29
> 20 Years of Experience in District	0.12
Graduate Degree	0.50
n	236

Notes: Missing data reduces the data available for some measures.

Table 3: The Predictive Validity of Self-Reported Character Skills on Education, Employment, Personal, and Civic Outcomes

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

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

Table 4: Correlations among State Tests, Complex Tasks and Socio-Emotional Measures

	State Math	State ELA	BAM Math	SAT9-OE Reading	Growth Mindset	Grit: Consistency	Grit: Perseverance
State Math	1.00						
State ELA	0.74	1.00					
BAM Math	0.66	0.58	1.00				
SAT9-OE Reading	0.43	0.49	0.53	1.00			
Growth Mindset	0.24	0.28	0.25	0.22	1.00		
Grit: Consistency	0.28	0.31	0.25	0.22	0.31	1.00	
Grit: Perseverance	0.19	0.21	0.15	0.17	0.03	0.23	1.00
Effort in Class	0.29	0.29	0.24	0.23	0.14	0.35	0.56

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

Table 5: Testing Post-Attrition Randomization Balance in Student Demographic and Prior Achievement across Teachers in the Same Randomization Block

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

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

Table 6: The Relationship between Student Characteristics and Randomly Assigned Teacher Characteristics Post-Attrition

	Teacher Value-Added			
	State Math	State ELA	BAM	SAT9-OE
Male	-0.002	0.000	-0.012	0.002
	(0.014)	(0.009)	(0.009)	(0.007)
Age	0.003	0.005	0.021	-0.011
	(0.014)	(0.010)	(0.015)	(0.012)
Gifted Status	0.024	-0.001	0.002	-0.007
	(0.020)	(0.013)	(0.014)	(0.010)
Special Education Status	0.009	0.003	0.013	0.003
	(0.008)	(0.006)	(0.011)	(0.012)
English Language Learner	-0.026	-0.018	-0.007	-0.013
	(0.014)	(0.011)	(0.018)	(0.014)
Free or Reduced Price Lunch	0.028*	0.002	0.001	0.017
	(0.013)	(0.008)	(0.015)	(0.018)
White	-0.014	-0.011	-0.004	-0.013
	(0.008)	(0.007)	(0.006)	(0.008)
African American	0.177	0.007	-0.007	0.015
	(0.009)	(0.009)	(0.010)	(0.012)
Hispanic	-0.160	-0.009	-0.001	0.000
	(0.010)	(0.010)	(0.012)	(0.011)
Asian	0.008	0.009	0.009	0.000
	(0.007)	(0.008)	(0.006)	(0.005)
State Math 2010	0.041	0.020	0.025	-0.017
	(0.036)	(0.021)	(0.026)	(0.028)
State ELA 2010	0.048	0.031	0.009	-0.018
	(0.030)	(0.022)	(0.027)	(0.025)
n	4092	4041	4076	4041
P-value of joint test of signifiance	0.000	0.168	0.376	0.548

Notes: *p<0.05. Each cell presents results from a separate regression of a given student characteristic on the value added of the teacher students were randomly assigned to by MET Project researchers. Joint hypothesis tests are estimated using the "seemingly unrelated regressions" or SURE model proposed by Zellner (1962).

Table 7: Model-based Restricted Maximium Likelihood Estimates of Teacher Effects on State Tests, Complex Tasks and Socio-Emotional Measures

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
State Math	4,075	0.159***	0.170***	0.140***	0.149***	0.121***
State ELA	4,074	0.141***	0.173***	0.143***	0.149***	0.125***
BAM Math	3,744	0.137***	0.168***	0.131***	0.148***	0.112**
SAT9-OE Reading	3,766	0.163***	0.180***	0.163***	0.174***	0.156***
Growth Mindset	3,551	0.201***	0.156**	0.138**	0.168***	0.158**
Grit: Consistency	3,473	0.089	0.089	0.075	0.097	0.101
Grit: Perseverance	3,473	0.151**	0.152**	0.140*	0.152**	0.142*
Effort in Class	3,435	0.159***	0.159**	0.173***	0.114*	0.141*
Survey Controls			Yes	Yes	Yes	Yes
Peer Controls		Yes		Yes		Yes
School FE		Yes				
Randomization Block FE			Yes	Yes	Yes	Yes

Notes: * p<0.05, ** p<0.01, *** p<0.001. Cells report the standard deviation of teacher effect estimates from separate regressions.

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

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	0.60***	1.00					
BAM Math	0.66***	0.36***	1.00				
SAT9-OE Reading	0.34***	0.25***	0.43***	1.00			
Growth Mindset	0.23***	0.19**	0.12	0.22***	1.00		
Grit: Consistency	0.18**	0.20**	0.10	-0.02	0.22***	1.00	
Grit: Perseverence	-0.06	-0.02	0.10	0.18**	-0.02	0.03	1.00
Effort in Class	0.07	0.08	0.14*	0.09	-0.05	0.06	0.61***

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. Classroom effects are derived using the model reported in Column 3 of Table 7.

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

	FFT	CLASS	Principal Ratings	Student Surveys
State Math	0.084	0.072	-0.169	0.001
State ELA	0.101	0.036	-0.110	0.067
BAM Math	0.115	0.058	-0.095	0.120
SAT9-OE Reading	0.120	0.074	-0.038	0.036
Growth Mindset	0.101	0.111	-0.157*	0.009
Grit: Consistency	0.039	0.025	-0.040	0.080
Grit: Perseverence	0.079	0.068	0.127	0.188**
Effort in Class	0.126	0.127	0.088	0.192**

Notes: *p<0.05; **p<0.01; ***p<0.001. Classroom effects are derived using the model reported in Column 3 of Table 7. n ranges from 191 (principal ratings) to 235 (FFT & CLASS).

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

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
State Math	4,075	0.142	0.127	0.094	0.097	0.107
State ELA	4,074	0.120	0.124	0.089	0.082	0.069
BAM Math	3,744	0.134	0.130	0.096	0.088	0.073
SAT9-OE Reading	3,766	0.160	0.146	0.114	0.117	0.103
Growth Mindset	3,551	0.214	0.157	0.108	0.136	0.107
Grit: Consistency	3,473	0.124	0.105	0.088	0.102	0.096
Grit: Perseverance	3,473	0.162	0.142	0.107	0.113	0.103
Effort in Class	3,435	0.178	0.152	0.121	0.137	0.161
Survey Controls			Yes	Yes	Yes	Yes
Peer Controls		Yes		Yes		Yes
School FE		Yes				
Randomization Block FE			Yes	Yes	Yes	Yes

Notes: Cells represent estimates from separate regressions. Statistical significance not calculated given estimates represent the average across shrunken and unshrunken estimates.

Table 11: Falification Tests of Teacher Effects

		Actual Teacher			Randomly Assigned Teacher (Intent to Treat)	
	n	(1)	(2)	(3)	(4)	(5)
Random Number	4,092	0.000	0.000	0.000	0.000	0.000
Male	4,092	0.000	0.000	0.000	0.000	0.000
Age	4,092	0.049	0.046	0.048	0.043	0.040
Free or Reduced Price Lunch	2,326	0.000	0.000	0.000	0.000	0.000
White	4,092	0.000	0.000	0.000	0.000	0.000
African American	4,092	0.030	0.022	0.025	0.000	0.000
Hispanic	4,092	0.028	0.032	0.029	0.029	0.024
Survey Controls			Yes	Yes	Yes	Yes
Peer Controls		Yes		Yes		Yes
School FE		Yes				
Randomization Block FE			Yes	Yes	Yes	Yes

Notes: Cell represent model-based restricted maximium likelihood estimates from separate regressions. No estimates are statistically significant. The sample size for Free or Reduced Price Lunch is limited because one participating district did not provide this information.

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

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

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

On-line 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)

On-line Appendix B

Measures used in the Educational Longitudinal Study analyses

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

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

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

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

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

Achievement Measures

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

Adult Outcome Measures

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

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

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

On-line Appendix C

We disattenuate raw correlations using the Spearman (1904) adjustment 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:

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

We calculate the reliably of our 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. We 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. We report these reliabilities in Table AC1 as well as disattenuated correlations across outcomes in Table AC2 below.

Table AC1 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

Table AC2: Disattenuated Correlations among State Tests, Complex Tasks and Socio-Emotional Measures

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	0.81	1.00					
BAM Math	0.81	0.73	1.00				
SAT9-OE Reading	0.49	0.56	0.68	1.00			
Growth Mindset	0.28	0.34	0.33	0.27	1.00		
Grit: Consistency	0.36	0.40	0.36	0.29	0.43	1.00	
Grit: Perseverence	0.24	0.27	0.21	0.22	0.04	0.34	1.00
Effort in Class	0.40	0.41	0.38	0.33	0.21	0.58	0.90

On-line Appendix D

We begin by examining the degree to which sampling error may attenuate our estimated correlation coefficients. We repeatedly re-estimate our correlation matrix within a sample of teachers with minimum class sizes ranging from six to 20 and plot the results in Figure AD1. These sample restrictions incrementally reduce our estimation sample from the original 227 teachers to 87 teachers at the maximum restriction of 20 students. Figure AD1 illustrates how increasing the minimum sample size for estimating teacher effect slightly increases the magnitude for most, but not all, pair-wise correlations. The average increase in estimated correlations across all 18 pairwise correlations is only 0.029.

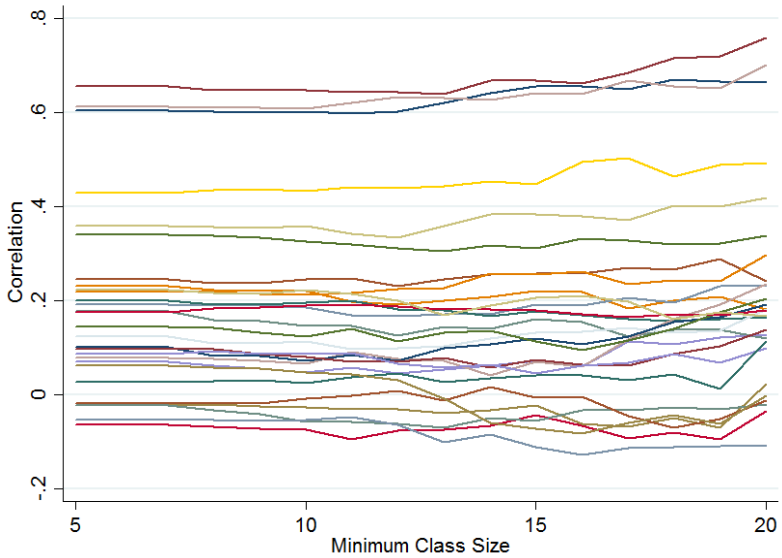


Figure AD1: Re-estimated correlations coefficients from Table 8 in restricted sample of successively larger minimum class size requirements

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

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

Table 7 provides our model-based ML estimate of σ_τ^2 for each outcome. We approximate $\sigma_{\varepsilon_j}^2$ as the average of the squared standard errors of post-hoc predicted BLUE teacher effects from our ML models ($\overline{SE_{\tau_j}^2}$). This approach can be seen as providing an upward bound estimate of the true correlations. For example, we find that correlations of approximately 0.60 and above are adjusted to be greater than 1, outside the possible range of correlation coefficients. We report our estimated reliabilities for each teacher effect in Table AD1 and disattenuated correlations in Table AD2 below.

Table AD1: Estimated Reliabilities of Teacher Effects	
State Math	0.566
State ELA	0.563
BAM Math	0.547
SAT9-OE Reading	0.554
Growth Mindset	0.533
Grit: Consistency	0.509
Grit: Perseverance	0.530
Effort in Class	0.543

Table AD2: Disattentuated Correlations among Teacher Effects on State Tests, Complex Tasks and Socio-Emotional Measures

	State Math	State ELA	BAM Math	SAT9 Reading	Growth Mindset	Grit: Consistency	Grit: Perseverence
State Math	1.00						
State ELA	1.00	1.00					
BAM Math	1.00	0.65	1.00				
SAT9-OE Reading	0.61	0.45	0.78	1.00			
Growth Mindset	0.42	0.35	0.22	0.40	1.00		
Grit: Consistency	0.34	0.37	0.19	-0.04	0.42	1.00	
Grit: Perseverence	-0.11	-0.04	0.19	0.33	-0.04	0.06	1.00
Effort in Class	0.13	0.14	0.26	0.16	-0.09	0.11	1.00

Notes: *p<0.05; **p<0.01; ***p<0.001. n = 227. Classroom effects are derived using the model reported in Column 3 of Table 5. Disattentuated estimates outside the range of correlation coefficients are set to 1.

Finally, we explore the degree to which unobserved student traits correlated with multiple outcomes results in an upward bias of our estimates. Past researchers have resolved this challenge by correlating teacher effects between different classes or years. We are unable to estimate teacher effects across classes given our focus on general elementary school teachers. The single administration of the survey questions capturing students' self-reported grit and growth mindset in the second year of the study also prevents us from comparing teacher effects across years for these outcomes. We attempt to better understand this issue by comparing correlations both within and between class sections for teachers who taught multiple classes in the first year of the MET Project. We use teacher effects provided in the MET Project data that are derived from standard value-added model (Kane & Cantrell, 2010). Estimates reported in Table AD3 are available for five different outcomes but can only be compared within-subject given the available data.

Table AD3: Teacher Effect Correlations within and between Sections

	n (Teachers)	Correlation Within Sections	Correlation Between Sections
State Math & BAM Math	433	0.38	0.14
State Math & Effort in Class	433	0.17	0.18
BAM Math & Effort in Class	433	0.19	0.16
State ELA & SAT9-OE Reading	497	0.33	0.21
State ELA & Effort in Class	497	0.12	0.09
SAT9-OE Reading & Effort in Class	497	0.14	0.13

Notes: Estimates derived using value-added estimates provided by the MET Project for subject-specialist teachers in Year 1 of the study. Disattenuated correlations are estimated by dividing the between-section correlations by the square root of the produce of the variance of each measure as described in footnote 9 of Kane and Cantrell (2010). The variance of each measure is estimated using the between-section correlations: State Math (.45), State ELA (.18), BAM Math (.29), SAT0-OE Reading (.46), and Effort in Class (.33).

We find that teacher effect correlations estimated from the same section of students are inflated relative to between-section teacher effects which eliminate the bias of common student shocks. The largest degree of inflation occurs for estimates among outcomes that are more highly correlated such as state tests and the supplemental open-ended assessments administered by the MET project. Smaller correlations between teacher effects on achievement measures and students' self-reported effort in class are only slightly inflated if at all. These findings suggest that sampling and measurement error likely dominate any upward bias in our estimates of the correlations between teacher effects on achievement measures and social-emotional competencies due to correlated errors.